Evidence-Based Uncertainty Sampling for Active Learning

Manali Sharma · Mustafa Bilgic

Abstract Active learning methods select informative instances to effectively learn a suitable classifier. Uncertainty sampling, a frequently utilized active learning strategy, selects instances about which the model is uncertain but it does not consider the reasons for why the model is uncertain. In this article, we present an evidence-based framework that can uncover the reasons for why a model is uncertain on a given instance. Using the evidence-based framework, we discuss two reasons for uncertainty of a model: a model can be uncertain about an instance because it has strong, but conflicting evidence for both classes or it can be uncertain because it does not have enough evidence for either class. Our empirical evaluations on several real-world datasets show that distinguishing between these two types of uncertainties has a drastic impact on the learning efficiency. We further provide empirical and analytical justifications as to why distinguishing between the two uncertainties matters.

Keywords Active learning · Uncertainty sampling · Classification

1 Introduction

Active learning methods interact with labelers, guiding them to the most informative instances to be annotated, to efficiently learn the correct classification function with minimum cost (e.g., time, money, and effort) (Settles, 2012). Many active learning strategies have been developed to date. Due to its simplicity, intuitiveness, and empirical success in many domains, uncertainty sampling (Lewis and Gale, 1994) is one of the most frequently utilized ones in the literature (e.g. (Bilgic et al, 2010), (Tong and Chang, 2001), and (Settles and Craven, 2008)). In a nutshell, it picks instances for which the model is most uncertain. Though it has many limitations, such as sensitivity to noise and outliers, it still works surprisingly well.

Traditional uncertainty sampling does not delve into the reasons for model's uncertainty on instances. In this article, we use the evidence-based framework to analyze why the model might be uncertain about an instance. Specifically, we focus on two types of uncertainties. In the first case, the

Manali Sharma
E-mail: msharm11@hawk.iit.edu
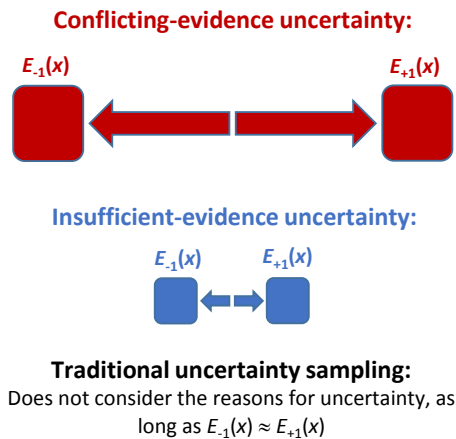
Mustafa Bilgic
E-mail: mbilgic@iit.edu
Illinois Institute of Technology
10 W 31st Street
Chicago, IL, 60616

model is uncertain due of presence of strong, but conflicting evidence for each class. We call this type of uncertainty as *conflicting-evidence uncertainty*. In the second case, the model is uncertain due to insufficient evidence for either class. We call this type of uncertainty as *insufficient-evidence uncertainty*.

For example, for a heart-disease diagnosis, the model can be uncertain because one lab test result strongly suggests presence of heart-disease, while another lab test result strongly suggests absence of heart-disease. In this case, the model is uncertain because of conflicting evidence for both classes. Another reason that the model can be uncertain is that none of the lab test results provide any conclusive evidence for presence or absence of heart-disease. In this case, the model is uncertain because of insufficient evidence for either class. Similarly, in a bag-of-words document classification task, the model can be uncertain because some terms in a document provide strong evidence for one class, while some other terms provide strong evidence for the other class, which makes the model uncertain due to conflicting evidence. On the other hand, the model can be uncertain because none of the terms provide conclusive evidence for either class, which represents model's uncertainty due to insufficient evidence. Fig. 1 depicts this phenomenon for binary classification.

Conflicting-evidence uncertainty:

$E_{-1}(x)$ $E_{+1}(x)$

Insufficient-evidence uncertainty:

$E_{-1}(x)$ $E_{+1}(x)$

Traditional uncertainty sampling:
Does not consider the reasons for uncertainty, as
long as $E_{-1}(x) \approx E_{+1}(x)$

Fig. 1 Conflicting-evidence vs. insufficient-evidence uncertainty. Conflicting-evidence uncertainty represents a model's uncertainty on an instance due to strong evidence for each class, whereas insufficient-evidence uncertainty represents a model's uncertainty on an instance due to insufficient evidence for each class. Traditional uncertainty sampling does not care about the reasons for uncertainty, and picks the most uncertain instance.

We provide a mathematical formalism to make a distinction between these two types of uncertainties. We introduce an evidence-based framework to capture the amount of evidence for each class provided by an instance, which facilitates distinguishing between these two types of uncertainties. Through empirical evaluations on several real-world datasets, we show that distinguishing between conflicting-evidence uncertainty and insufficient-evidence uncertainty makes a huge difference to the performance of active learning. We show that conflicting-evidence uncertainty provides the most benefit for learning, drastically outperforming both traditional uncertainty sampling and insufficient-evidence uncertainty sampling.

This article builds upon our earlier work (Sharma and Bilgic, 2013). Our contributions in (Sharma and Bilgic, 2013) were:

– We introduced an evidence-based framework to distinguish between the two types of uncertainties, namely conflicting-evidence uncertainty and insufficient-evidence uncertainty.

- We empirically evaluated our methods on several real-world datasets and showed that distinguishing between the reasons for uncertainty is useful to improve active learning.
- We provided formulation of evidence for naïve Bayes, logistic regression, and support vector machines, and extended the evidence-based framework for multi-class classification.

Our additional contributions in this article include:

- We provide empirical and analytical justifications as to why distinguishing between different types of uncertainties matters. Specifically, we show that the instances that are uncertain due to conflicting evidence have lower density in the labeled set, compared to instances that are uncertain due to insufficient evidence. That is, there is less support in the training data for the perceived conflict than for the insufficiency of the evidence.
- We provide empirical results showing that the model's variance on uncertain instances with conflicting evidence is higher than the model's variance on uncertain instances that have insufficient evidence. This is partly because the parameters that lead to conflict are not supported by a lot of labeled data and therefore they lead to a higher variance.
- We compare performance of conflicting-evidence and insufficient-evidence uncertainties to query-by-committee strategy, which is the most similar active learning baseline for our study, because query-by-committee also chooses instances on which model has high prediction variance.
- We present results of a user study that examines users' performance on conflicting-evidence and insufficient-evidence cases.
- We test sensitivity of the proposed approaches to hyperparameters and provide more in-depth evaluation of rank of uncertain instances selected by different methods.
- We provide formulation of evidence-based framework for kernel-based support vector machines where kernels do not need to be linear.

The rest of the article is organized as follows. In Sect. 2, we provide background on active learning and uncertainty sampling, and provide formulation for the evidence-based framework. In Sect. 3, we provide experimental details and results comparing conflicting-evidence uncertainty and insufficient-evidence uncertainty to the traditional uncertainty sampling. In Sect. 4, we present results of a user study that examines the users' performance while labeling instances selected by the two types of uncertainties. In Sect. 5, we present empirical and analytical justifications as to why distinguishing between conflicting versus insufficient evidence cases matters. In Sect. 6, we extend the formulation for the evidence-based framework to other classifiers and multi-class classification. Finally, we conclude and present future work in Sect. 7.

2 Background and Problem Formulation

In this section, we first provide background on active learning and uncertainty sampling. Then we explain active learning and uncertainty sampling in the context of classification. Then we provide the formulation for the evidence-based framework for naïve Bayes.

Many active learning methods have been developed in the past two decades. A number of approaches have been proposed to select informative instances for labeling, e.g. selecting uncertain instances (Lewis and Gale, 1994), choosing instances for which a committee of learners disagree (Seung et al, 1992), choosing representative instances (Xu et al, 2003), selecting more informative data that optimizes expected gain (MacKay, 1992), selecting examples that minimize the expected error of the model ((Roy and McCallum, 2001), (Yu et al, 2006), (Gu et al, 2012), (Gu et al, 2014)), and selecting instances that

minimize the bias of the learner (Cohn, 1997) or minimize variance of the learner (Cohn et al, 1996). We refer the reader to (Settles, 2012) for a survey of active learning methods.

Arguably, the most frequently utilized active learning strategy is uncertainty sampling. [1]It is often used as a baseline for comparing other active learning methods and has been shown to work successfully in a variety of domains. Example domains include text classification (Lewis and Gale, 1994),(Bilgic et al, 2010),(Xu et al, 2003),(Hoi et al, 2006a), natural language processing (Thompson et al, 1999), email spam filtering (Sculley, 2007),(Segal et al, 2006), image retrieval (Tong and Chang, 2001), medical image classification (Hoi et al, 2006b), robotics (Chao et al, 2010), information retrieval (Zhang and Chen, 2002), dual supervision (Sindhwani et al, 2009), and sequence labeling (Settles and Craven, 2008), among many others.

Even though uncertainty sampling is frequently utilized, it is known to be susceptible to noise and outliers (Roy and McCallum, 2001). A number of approaches have been proposed to make it more robust. For example, (Settles and Craven, 2008) weights the uncertainty of an instance by its density to avoid outliers, where density of the instance is defined as average similarity to other instances. (Zhu et al, 2008) used a K-Nearest-Neighbor-based density measure to determine whether an unlabeled instance is an outlier. (Xu et al, 2003) and (Donmez et al, 2007) proposed a hybrid approach to combine representative sampling and uncertainty sampling. Other approaches used the cluster structure of the domain to choose more representative examples (Nguyen and Smeulders, 2004), (Bilgic et al, 2010). (Senge et al, 2014) presented an approach to distinguish between aleatoric and epistemic uncertainties using possibility theory, in which uncertainty is modeled in terms of two measures, namely possibility and necessity. The aleatoric uncertainty results due to variability in the outcome of an experiment due to inherently random effects, and epistemic uncertainty is caused by lack of knowledge. While epistemic uncertainty can be reduced by gathering more information, aleatoric uncertainty cannot be reduced further.

Our work is orthogonal to these approaches. We are not providing yet another alternative approach to improve uncertainty sampling, but instead we are highlighting that distinguishing between the two types of uncertainties (conflicting-evidence vs. insufficient-evidence) has a big impact on active learning. One can imagine combining uncertainty sampling, density weighting, and conflicting-evidence uncertainty methods because they are not mutually exclusive.

Next, we explain active learning and uncertainty sampling in detail and introduce the notations that will be used throughout the article.

2.1 Active Learning

Let the uppercase X denote the random variable representing an instance and the lowercase x represent a particular instantiation of X. Each instance is described as a vector of f attributes $X \triangleq \langle X_1, X_2, \cdots, X_f \rangle$. Similarly, let the uppercase Y represent the class variable of the instance and let the lowercase y represent a particular instantiation of Y. Each X_i can be real-valued or discrete whereas Y is discrete; in this article, we focus on the binary case, where $Y \in \{-1, +1\}$. In the pool-based active learning setup, we are given a small set of instances whose labels are known: $\mathcal{L} = \{\langle x^{(i)}, y^{(i)} \rangle\}$, and a much larger collection of unlabeled instances whose labels are unknown: $\mathcal{U} = \{\langle x^{(i)}, ? \rangle\}$.

A pool-based greedy active learning algorithm iteratively selects an informative instance $\langle x^*, ? \rangle \in \mathcal{U}$ to obtain its label y^* from an expert, and incorporates the new labeled instance $\langle x^*, y^* \rangle$ into $\mathcal{L}$. The informative instance, $\langle x^*, ? \rangle$, is selected by computing *utility* of the unlabeled instances in $\mathcal{U}$, where

---

[1] 1,507 citations on Google Scholar on April 4th, 2016

utility can be classifier uncertainty (Lewis and Gale, 1994), committee disagreement (Seung et al, 1992), expected reduction in error (Roy and McCallum, 2001), etc. This process continues until a stopping criterion is met, usually until a given budget, B, is exhausted. Algorithm 1 describes this process more formally. The goal of active learning is to learn the correct classification function $\theta : X \to Y$ by carefully choosing which instances are labeled, subject to budgetary constraints.

Algorithm 1 Pool-Based Active Learning

1: **Input:** $\mathcal{U}$ - unlabeled data, $\mathcal{L}$ - labeled data, θ - classification model, B - budget
2: **repeat**
3:    **for all** $\langle x^{(i)}, ? \rangle \in \mathcal{U}$ **do**
4:       compute $utility(x^{(i)}, \theta)$
5:    **end for**
6:    pick highest utility x^* and query its label
7:    $\mathcal{L} \leftarrow \mathcal{L} \cup \{\langle x^*, y^* \rangle\}$
8:    $\mathcal{U} \leftarrow \mathcal{U} \setminus \{\langle x^*, y^* \rangle\}$
9:    Train θ on $\mathcal{L}$
10:   $B = B - 1$
11: **until** $B == 0$

2.2 Uncertainty Sampling

Uncertainty sampling selects instances for which the current model is most uncertain how to label (Lewis and Gale, 1994). These instances correspond to the ones that lie close to the decision boundary of the current model. Uncertainty of an underlying model can be measured in several ways. One approach is to use conditional entropy:

$$x^* = \underset{x^{(i)} \in \mathcal{U}}{\arg\max} - \sum_{y \in Y} P_\theta(y|x^{(i)}) \log\left(P_\theta(y|x^{(i)})\right) \tag{1}$$

where $P_\theta(y|x^{(i)})$ is the probability that instance $x^{(i)}$ has label y. Another approach is to use maximum conditional:

$$x^* = \underset{x^{(i)} \in \mathcal{U}}{\arg\max} \left(1 - \max_{y \in Y} P_\theta(y|x^{(i)})\right) \tag{2}$$

The last approach we discuss uses margin of confidence:

$$x^* = \underset{x^{(i)} \in \mathcal{U}}{\arg\min} \left(P_\theta(y_m|x^{(i)}) - P_\theta(y_n|x^{(i)})\right) \tag{3}$$

where, y_m is the most likely label and y_n is the next likely label for $x^{(i)}$. More formally, $y_m = \underset{y \in Y}{\arg\max} P_\theta(y|x^{(i)})$ and $y_n = \underset{y \in Y \setminus y_m}{\arg\max} P_\theta(y|x^{(i)})$.

When the task is binary classification, that is when $Y \in \{+1, -1\}$, all three uncertainty approaches (Eq. 1, Eq. 2 and Eq. 3) rank instances in the same order and prefer the same uncertain instance, i.e. the instance for which $P_\theta(+1|x^{(i)}) = P_\theta(-1|x^{(i)}) = 0.5$. In this article, we distinguish between the two types of uncertainties that we define next.

2.3 Problem Formulation

In this section, we define *evidence* that an attribute value provides for a class in the evidence-based framework. The evidence, in its most general form, is the amount of contribution that an attribute value provides to the prediction of belonging to a particular class. Each classifier computes the prediction for a test instance differently, and hence the evidence that an attribute value of an instance provides for a class depends on the classifier. In this section, we provide the formalism of evidence using naïve Bayes classifier. The formalism of evidence for logistic regression and support vector machines is provided later in Sect. 6.

2.3.1 Evidence using naïve Bayes

A naïve Bayes classifier uses the Bayes rule to compute $P(Y|X)$ and assumes that the attributes X_j are conditionally independent given Y:

$$P(Y|x) = P(Y|x_1, x_2, \cdots, x_f) = \frac{P(Y)\prod\limits_{x_j} P(x_j|Y)}{P(x_1, x_2, \cdots, x_f)} \tag{4}$$

The instance x can be classified based on the ratio of $\frac{P(+1|x)}{P(-1|x)}$:

$$Y = \begin{cases} +1 & \text{if } \left(\frac{P(+1)}{P(-1)} \prod\limits_{x_j} \frac{P(x_j|+1)}{P(x_j|-1)} \right) > 1 \\ -1 & \text{otherwise} \end{cases} \tag{5}$$

From Eq. 5, it follows that the attribute value $x_j^{(i)}$ of the instance $x^{(i)}$ provides evidence for the positive class if $\frac{P(x_j^{(i)}|+1)}{P(x_j^{(i)}|-1)} > 1$, and it provides evidence for the negative class otherwise.

Note that it does not make sense to talk about the evidence the attribute X_j itself provides. Rather, the particular instantiation x_j provides evidence for one class or the other (or for none of the classes). For example, the cholesterol test itself does not provide evidence for presence or absence of heart-disease; rather, the outcome of the cholesterol test (e.g., high or low) provides the evidence for presence/absence of heart-disease. Hence, we define the evidence at the instance level, rather than the variable level.

Let $\mathcal{P}_{x^{(i)}}$ and $\mathcal{N}_{x^{(i)}}$ be two sets, such that $\mathcal{P}_{x^{(i)}}$ contains the attribute values of the instance $x^{(i)}$ that provide evidence for the positive class and $\mathcal{N}_{x^{(i)}}$ contains the attribute values of the instance $x^{(i)}$ that provide evidence for the negative class:

$$\mathcal{P}_{x^{(i)}} \triangleq \left\{ x_j^{(i)} \mid \frac{P(x_j^{(i)}|+1)}{P(x_j^{(i)}|-1)} > 1 \right\}$$

$$\mathcal{N}_{x^{(i)}} \triangleq \left\{ x_k^{(i)} \mid \frac{P(x_k^{(i)}|-1)}{P(x_k^{(i)}|+1)} > 1 \right\}$$

Note that in these definitions, the numerator for $\mathcal{P}_{x^{(i)}}$ is $P(x_j^{(i)}|+1)$ and numerator for $\mathcal{N}_{x^{(i)}}$ is $P(x_k^{(i)}|-1)$.

The total evidence for the instance $x^{(i)}$ to belong to the positive class is:

$$E_{+1}(x^{(i)}) = \prod_{x_j^{(i)} \in \mathcal{P}_{x^{(i)}}} \frac{P(x_j^{(i)}|+1)}{P(x_j^{(i)}|-1)} \tag{6}$$

and, the total evidence for the instance $x^{(i)}$ to belong to the negative class is:

$$E_{-1}(x^{(i)}) = \prod_{x_k^{(i)} \in \mathcal{N}_{x^{(i)}}} \frac{P(x_k^{(i)}|-1)}{P(x_k^{(i)}|+1)} \tag{7}$$

With these definitions, we can rewrite the classification rule for naïve Bayes as:

$$Y = \begin{cases} +1 & \text{if } \left(\frac{P(+1)}{P(-1)} \frac{E_{+1}(x^{(i)})}{E_{-1}(x^{(i)})} \right) > 1 \\ -1 & \text{otherwise} \end{cases} \tag{8}$$

2.3.2 Conflicting-Evidence vs. Insufficient-Evidence Uncertainty

In this article, we investigate whether the evidence-based framework provides a useful criteria to distinguish between the uncertain instances and whether such an approach leads to more or less effective active learning.

Traditional uncertainty sampling picks the most uncertain instance, $x^{(i)}$, for which $E_{+1}(x^{(i)}) \approx E_{-1}(x^{(i)})$, regardless of the magnitudes of $E_{+1}(x^{(i)})$ and $E_{-1}(x^{(i)})$. In this article, we analyze if the magnitudes of $E_{+1}(x^{(i)})$ and $E_{-1}(x^{(i)})$ have an impact on learning when $E_{+1}(x^{(i)}) \approx E_{-1}(x^{(i)})$. Specifically, we consider two cases:

- The model is uncertain because of strong, but conflicting evidence for both classes. This represents the case when both $E_{+1}(x^{(i)})$ and $E_{-1}(x^{(i)})$ are equal and large.
- The model is uncertain because of insufficient evidence for either class. This represents the case when both $E_{+1}(x^{(i)})$ and $E_{-1}(x^{(i)})$ are equal and small.

When $E_{+1}(x^{(i)})$ and $E_{-1}(x^{(i)})$ are equal, there are a number of choices to mathematically determine if both $E_{+1}(x^{(i)})$ and $E_{-1}(x^{(i)})$ are small or large by ranking all the uncertain instances according to one of the Equations 9, 10, 11, or 12.

$$\underset{x^{(i)} \in \mathcal{U}}{\arg\max} \, E_{+1}(x^{(i)}) \times E_{-1}(x^{(i)}) \tag{9}$$

$$\underset{x^{(i)} \in \mathcal{U}}{\arg\max} \, E_{+1}(x^{(i)}) \tag{10}$$

$$\underset{x^{(i)} \in \mathcal{U}}{\arg\max} \, E_{-1}(x^{(i)}) \tag{11}$$

$$\underset{x^{(i)} \in \mathcal{U}}{\arg\max} \, \min(E_{+1}(x^{(i)}), E_{-1}(x^{(i)})) \tag{12}$$

Note that when $E_{+1}(x^{(i)})$ and $E_{-1}(x^{(i)})$ are equal, Equations 9, 10, 11, and 12 will all provide the same ranking for uncertain instances, and it does not matter which one of these functions is chosen to rank the uncertain instances based on evidence. [2] In Sect. 3.2, we present the results using multiplication of the evidence for each class, i.e. according to Eq. 9.

Regardless of whether we want to maximize or minimize $E_{+1}(x) \times E_{-1}(x)$, we want to guarantee that the underlying model is uncertain about the chosen instance. To achieve uncertainty, we first rank the instances $x^{(i)} \in \mathcal{U}$ in decreasing order of their uncertainty score (measured by Eq. 1), and work

---

[2] In practice, however, $E_{+1}(x^{(i)})$ and $E_{-1}(x^{(i)})$ might not be exactly equal to each other for all uncertain instances, and hence the ranking of uncertain instances based on evidence according to Equations 9, 10, 11, and 12 may be different.

with the top t instances, where t is a hyper-parameter. Formally, let $\mathcal{S}$ be the set of top t uncertain instances. Conflicting-evidence uncertainty will prefer instances where both $E_{+1}(x^{(i)})$ and $E_{-1}(x^{(i)})$ are large:

$$x^* = \arg\max_{x^{(i)} \in \mathcal{S}} E_{+1}(x^{(i)}) \times E_{-1}(x^{(i)}) \tag{13}$$

and, insufficient-evidence uncertainty will prefer instances where both $E_{+1}(x^{(i)})$ and $E_{-1}(x^{(i)})$ are small:

$$x^* = \arg\min_{x^{(i)} \in \mathcal{S}} E_{+1}(x^{(i)}) \times E_{-1}(x^{(i)}) \tag{14}$$

3 Experimental Methodology and Results

We designed our experiments to test whether distinguishing between the conflicting-evidence and insufficient-evidence uncertain instances makes a difference to the performance of active learner. We experimented with the following approaches:

1. *Random Sampling* (RND): This is a common baseline for active learning, in which instances are picked at random from the set of candidate unlabeled instances.
2. *Uncertainty Sampling - 1st* (UNC-1): This is the traditional uncertainty sampling method that picks the instance for which the underlying model is most uncertain, as defined in Sect. 2.2.
3. *Conflicting-Evidence Uncertainty* (UNC-CE): Among the top t uncertain instances, this method picks the instance for which the model is uncertain due to conflicting evidence (as defined in Eq. 13).
4. *Insufficient-Evidence Uncertainty* (UNC-IE): Among the top t uncertain instances, this method picks the instance for which the model is uncertain due to insufficient evidence (as defined in Eq. 14).
5. *Uncertainty Sampling - tth* (UNC-t): Among the top t uncertain instances, this method picks the tth most uncertain instance. UNC-CE and UNC-IE methods pick one uncertain instance from the top t uncertain instances according to the amount of evidence they provide. If UNC-CE and/or UNC-IE are better than UNC-1, then this result would suggest that different types of uncertainties matter. Similarly, if UNC-CE and/or UNC-IE are worse than UNC-t, then this result would also suggest that different types of uncertainties matter.

We experimented with eight publicly available datasets. We chose four medium-imbalanced (minority class% > 10%) and four highly-imbalanced (minority class% ≤ 10%) datasets. The datasets include four active learning challenge datasets (Guyon et al, 2011) (Ibn Sina, Nova, Zebra, and Hiva), and four additional datasets: LetterO (Frey and Slate, 1991), Calif. Housing (Pace and Barry, 1997), Spambase (Frank and Asuncion, 2010), and a thyroid disease dataset, Sick (Frank and Asuncion, 2010). The description of these datasets is provided in Table 1. We evaluated the five methods using three performance measures: AUC, accuracy, and F1. We computed F1 as a harmonic mean of precision and recall using the minority class as positive labels. We computed AUC for all the datasets, accuracy for only medium-imbalanced datasets (the top four in Table 1) and F1 for only highly-imbalanced datasets (bottom four in Table 1).

3.1 Parameters and Repeatability

We performed five-fold cross validation and the train split was treated as the unlabeled set, $\mathcal{U}$. 10 instances (five from each class) were chosen randomly and used as the initially labeled set, $\mathcal{L}$. For each fold, the experiment was repeated five times using different sets of 10 randomly chosen instances

Table 1 Description of the datasets: the domain, number of instances, number of features, types of features, and the percentage of minority class in the datasets. The datasets are sorted in increasing order of class imbalance.

Dataset	Domain	# of Instances	# of Features	Types of Features	Min. %
Spambase	Email classification	4,601	57	Numeric	39.4%
Ibn Sina	Handwriting recognition	20,722	92	Numeric	37.8%
Calif. Housing	Social	20,640	8	Numeric	29%
Nova	Text processing	19466	16969	Binary	28.4%
Sick	Medical	3,772	29	Numeric + Binary	6.1%
Zebra	Embryology	61,488	154	Numeric	4.6%
LetterO	Letter recognition	20,000	16	Numeric	4%
Hiva	Chemoinformatics	42,678	1617	Binary	3.5%

at bootstrap. At each iteration of active learning, the methods pick only one instance to be labeled. The budget, B, in Algorithm 1 was set to 500 instances. UNC-CE and UNC-IE operate within top t uncertain instances, as described in Sect. 2.3.2. We experimented with $t = 5$, 10, and 20. We evaluated each method using a naïve Bayes classifier with Laplace smoothing. To speed up the experiments, at each iteration we computed uncertainty over a set of randomly sub-sampled 250 instances, which is a common practice in active learning. The source code for evidence-based framework for naïve Bayes is available at http://www.cs.iit.edu/~ml/code/.
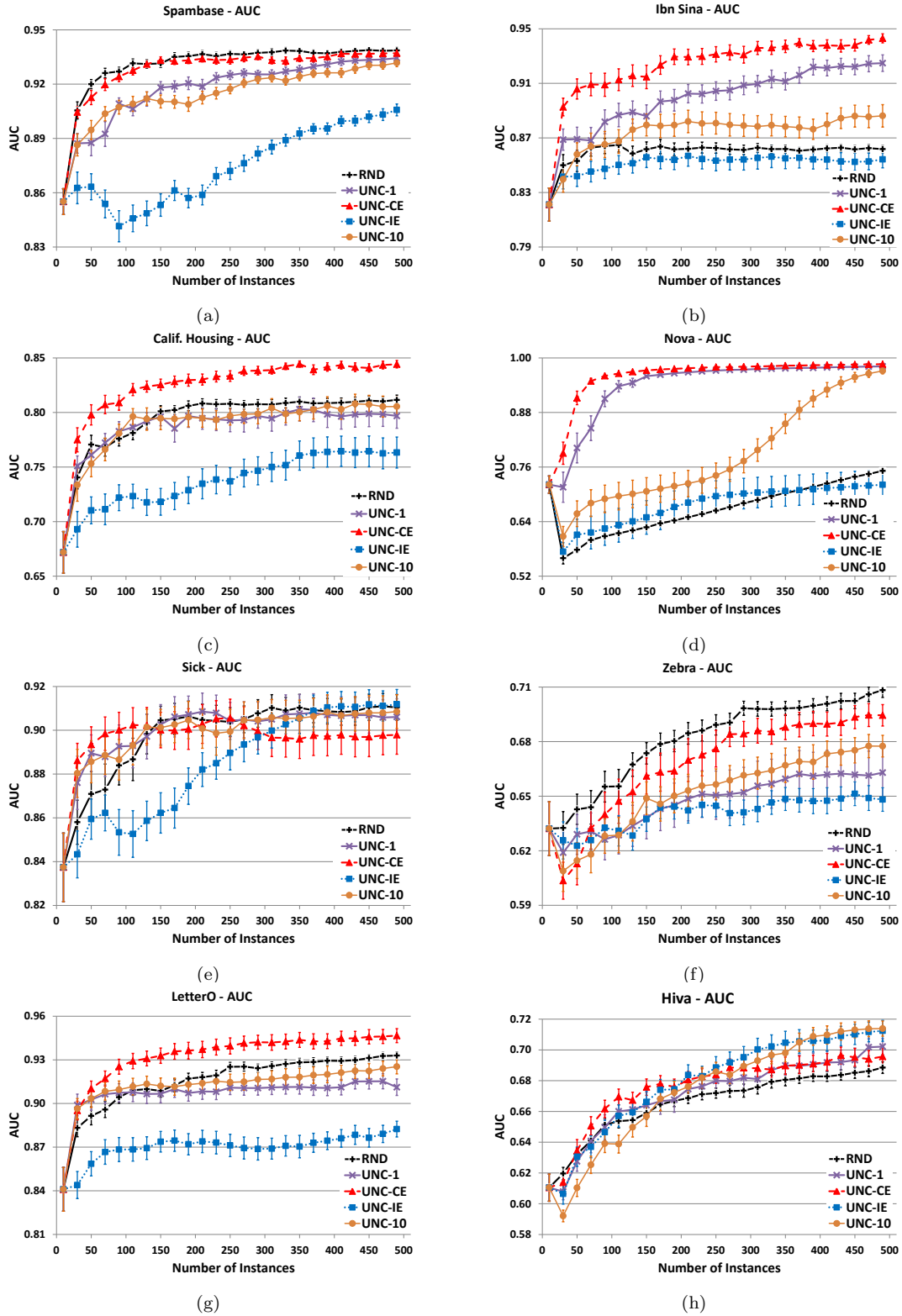
3.2 Results

In this section, we present the results for the five strategies presented in the beginning of Sect. 3 and show that distinguishing between the two types of uncertainties (conflicting-evidence uncertainty and insufficient-evidence uncertainty) makes a huge difference to the performance of active learning. We compare UNC-CE and UNC-IE strategies with both UNC-1 and UNC-t strategies. We use RND as a reference for the UNC-1 strategy.

We present the learning curves for RND, UNC-1, UNC-CE, UNC-IE, and UNC-t using $t=10$. The learning curves for UNC-CE, UNC-IE, and UNC-t with $t = 5$ and $t = 20$ are similar and are omitted to avoid redundancy. We present the AUC results in Fig. 2, and accuracy and F1 results in Fig. 3; these figures show the mean performance and $\pm$ standard error. As these figures show, distinguishing between conflicting-evidence and insufficient-evidence uncertain instances has a huge impact on active learning for all datasets and performance measures. UNC-CE wins over UNC-1 on most datasets and measures, whereas UNC-IE loses to UNC-1 on most datasets and measures.

Next, we present the results of t-tests comparing UNC-CE and UNC-IE to UNC-1 and UNC-t. Table 2 presents a summary of pairwise one-tailed t-tests results under significance level of 0.05, where the pairs are learning curves of the methods. If a method is statistically significantly better than the baseline, it is a Win (W), if it is statistically significantly worse than the baseline, it is a Loss (L), otherwise it a Tie (T), meaning the differences are not statistically significant. Note that for each method, the total counts of 'W', 'T' and 'L' should add up to 8 for AUC, 4 for accuracy, and 4 for F1.

Table 2 presents a summary of 'Win/Tie/Loss' counts of UNC-CE and UNC-IE with $t = 5$, $t = 10$, and $t = 20$ compared to UNC-1 baseline. With respect to UNC-1, there is a clear difference between UNC-CE and UNC-IE. Our results show that UNC-CE statistically significantly wins over UNC-1 on at least 6 out

Fig. 2 AUC results for all eight datasets. `UNC-CE` significantly outperforms `UNC-1` on seven out of eight datasets ((a), (b), (c), (d), (f), (g), and (h)) and loses on Sick dataset (e). `UNC-IE` loses to `UNC-1` on seven out of eight datasets ((a), (b), (c), (d), (e), (f), and (g), and wins on Hiva dataset (h).

(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)
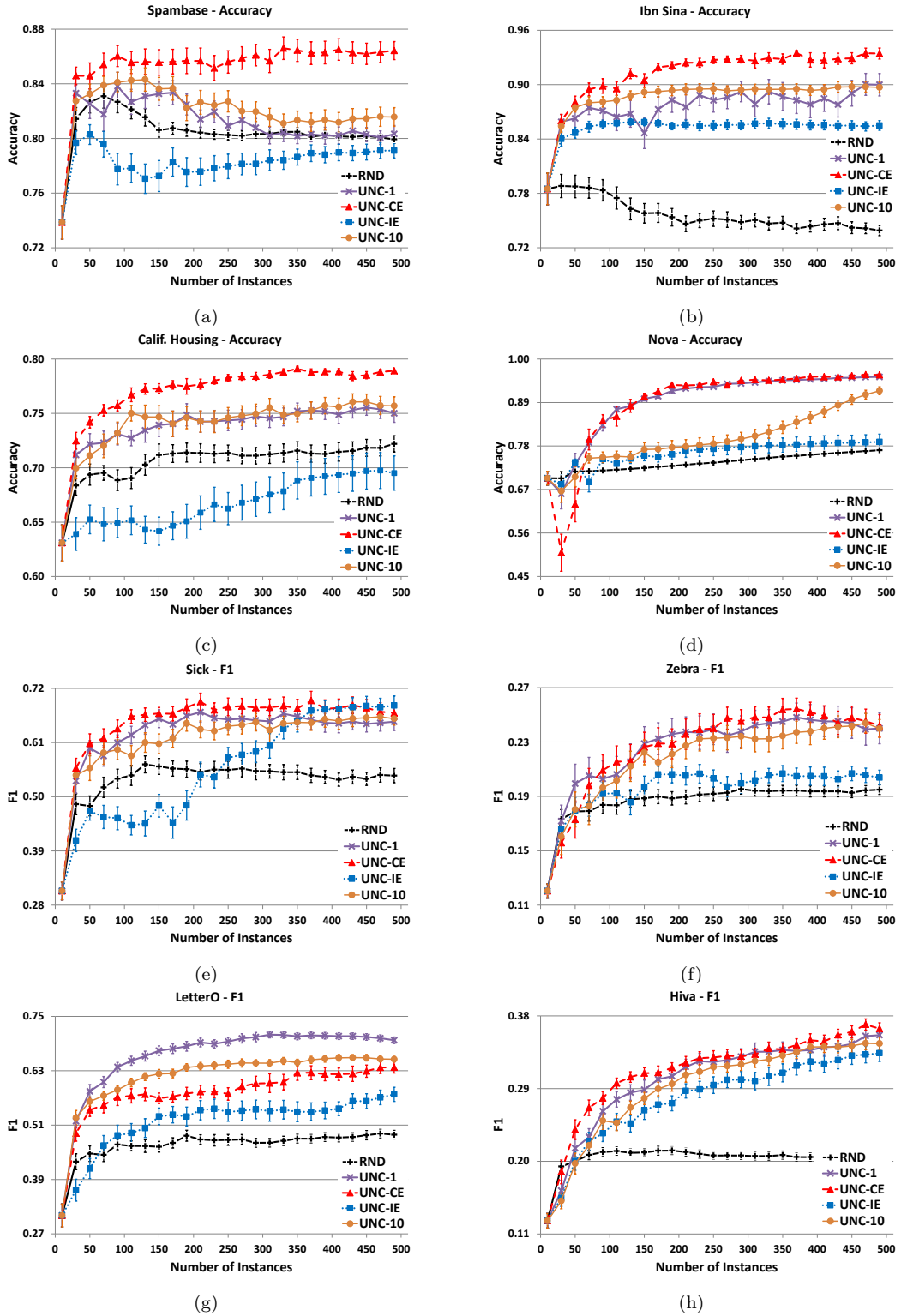
Fig. 3 Accuracy results for four medium-imbalanced datasets (Spambase, Ibn Sina, Calif. Housing and Nova). UNC-CE outperforms UNC-1 on three datasets ((a), (b) and (c)) and loses on Nova (d). UNC-IE loses to UNC-1 on all four datasets. F1 results for four relatively skewed datasets (Sick, Zebra, LetterO and Hiva). UNC-CE outperforms UNC-1 significantly on three datasets ((e), (f) and (h)), and loses on one (g). UNC-IE loses to UNC-1 on all four datasets.

of 8 datasets on AUC and loses on at most two datasets, whereas `UNC-IE` loses to `UNC-1` on 7 out of 8 datasets on AUC. On accuracy, `UNC-CE` wins over `UNC-1` on at least 3 out of 4 datasets, and loses on one dataset (Nova), whereas `UNC-IE` loses to `UNC-1` on all 4 datasets. On F1, `UNC-CE` wins on at least 2 out of 4 datasets and loses on one dataset (LetterO), whereas `UNC-IE` loses to `UNC-1` on all 4 datasets.

Table 2 `UNC-CE` and `UNC-IE` with $t = 5$, $t = 10$, and $t = 20$ versus `UNC-1`. Number of datasets on which `UNC-CE` and `UNC-IE` significantly Win (W), Tie (T), or Lose (L) compared to `UNC-1` baseline.

`UNC-1` **Baseline**	**AUC**	**ACCU**	**F1**
Method	**W/T/L**	**W/T/L**	**W/T/L**
`UNC-CE` with t=5	**7**/0/1	**4**/0/0	**3**/0/1
`UNC-CE` with t=10	**7**/0/1	**3**/0/1	**3**/0/1
`UNC-CE` with t=20	**6**/0/2	**3**/0/1	**2**/1/1
`UNC-IE` with t=5	1/0/**7**	0/0/**4**	0/0/**4**
`UNC-IE` with t=10	1/0/**7**	0/0/**4**	0/0/**4**
`UNC-IE` with t=20	1/0/**7**	0/0/**4**	0/0/**4**

`UNC-CE` not only wins over `UNC-1` for all performance measures, but is also quite efficient in saving the number of labeled instances required to achieve a target performance. For example, in order to achieve a target AUC of 80% for Calif. Housing dataset, `UNC-1` required 199 labeled instances, `UNC-CE` with $t = 10$ required only 59 labeled instances (70.4% savings in the number of labels), and `UNC-IE` could not achieve this target AUC even with 500 labeled instances. As another example, in order to achieve a target accuracy of 90% on Ibn Sina dataset, `UNC-1` required 344 labeled instances, `UNC-CE` with $t = 10$ required only 71 labeled instances (79.4% savings in the number of labels), and `UNC-IE` with $t = 10$ could not achieve this target accuracy even with 500 labeled instances. On Sick dataset, in order to achieve a target F1 of 65%, `UNC-1` required 127 labeled instances, `UNC-CE` with $t = 10$ required only 100 labeled instances (21.3% savings in the number of labels), and `UNC-IE` with $t = 10$ required 345 labeled instances to achieve this target F1.

Next, we compared `UNC-CE` and `UNC-IE` to `UNC-t` with $t = 5$, $t = 10$, and $t = 20$. Table 3 presents the 'Win/Tie/Loss' results using `UNC-t` as the baseline. We observe that `UNC-CE` significantly outperforms `UNC-t` on almost all datasets, which is not surprising because even a strategy that selects instances randomly from the top t uncertain instances has the potential to outperform `UNC-t`. However, it is surprising to observe that `UNC-IE` performs statistically significantly worse than `UNC-t` for almost all datasets and measures. Selecting uncertain instances that have insufficient evidence often performs worse than selecting the least uncertain instance among the top t uncertain instances. Note that `UNC-1`, `UNC-CE`, `UNC-IE`, and `UNC-t` do not have much flexibility in choosing between uncertain instances; that is they all work within the top t uncertain instances, and yet `UNC-IE` performs much worse than both `UNC-1` and `UNC-t`, whereas `UNC-CE` performs much better than both `UNC-1` and `UNC-t`.

`UNC-CE` clearly stands out as a winner strategy, whereas `UNC-IE` is clearly the worst performing uncertainty strategy. `UNC-CE` improves over `UNC-1` on almost all datasets and measures, whereas `UNC-IE` loses to `UNC-1` on almost all datasets and measures. This result is surprising because one would not expect such a huge difference between `UNC-CE` and `UNC-IE` strategies. After all, `UNC-CE` strategy picks an uncertain instance that has large evidence for both classes and hence, intuitively, labeling such instances is focused on correcting the mistakes of the learner. On the other hand, `UNC-IE` strategy picks an uncertain instance that has little evidence for both classes and hence, intuitively, labeling such instances is focused on teaching new things to the learner. Both types of uncertainties are expected

Table 3 `UNC-CE` and `UNC-IE` versus `UNC-t` with $t = 5$, $t = 10$, and $t = 20$. Number of datasets on which `UNC-CE` and `UNC-IE` significantly Win (W), Tie (T), or Lose (L) compared to `UNC-t` baseline.

`UNC-t` Baseline	AUC	ACCU	F1
Method	W/T/L	W/T/L	W/T/L
`UNC-CE` with $t = 5$	**6**/0/2	**4**/0/0	**3**/0/1
`UNC-CE` with $t = 10$	**7**/0/1	**4**/0/0	**3**/0/1
`UNC-CE` with $t = 20$	**6**/1/1	**4**/0/0	**3**/0/1
`UNC-IE` with $t = 5$	0/0/**8**	0/0/**4**	0/0/**4**
`UNC-IE` with $t = 10$	1/0/**7**	0/0/**4**	0/0/**4**
`UNC-IE` with $t = 20$	1/0/**7**	0/0/**4**	0/0/**4**

to be important for improving the model. We provide analytical and empirical justifications as to why `UNC-CE` outperforms `UNC-IE` in Sect. 5.

Next, we present a comparison of the ranks of the uncertain instances selected by `UNC-CE` and `UNC-IE`. Note that `UNC-1` will always pick the top most uncertain instance, and hence would select rank 1 uncertain instance. `UNC-t` on the other hand would always select rank t uncertain instance. `UNC-CE` and `UNC-IE` work within the top t uncertain instances and select rank u uncertain instance, where u is between 1 and t. Table 4 presents the mean rank of uncertain instances selected by `UNC-CE` and `UNC-IE` with $t = 10$ for all datasets. Fig. 4 presents histograms for all eight datasets, showing the ranks of uncertain instances selected by `UNC-CE` and `UNC-IE` with $t = 10$. The histograms with $t = 5$ and $t = 20$ have similar trends and are omitted to avoid redundancy. These histograms show that `UNC-CE` and `UNC-IE` choose a variety of ranks of uncertain instances for most datasets and hence the differences between `UNC-1`, `UNC-t`, `UNC-CE`, and `UNC-IE` do not stem from the rank of uncertain instances but rather, they are due to the information content of the different instances chosen by each method.

Table 4 The mean rank of uncertain instances selected by `UNC-CE` and `UNC-IE` for the eight datasets over various iterations of learning and 25 trials.

Dataset	`UNC-CE`		`UNC-IE`	
	Mean	Std. Dev	Mean	Std. Dev
Spambase	5.50	2.99	6.14	2.97
Ibn Sina	5.09	3.15	7.09	2.53
Calif. Housing	5.57	2.87	5.58	2.89
Nova	5.42	2.86	6.02	2.76
Sick	7.16	2.73	6.03	2.85
Zebra	5.92	2.88	6.03	2.85
LetterO	7.18	2.73	5.39	2.88
Hiva	5.27	2.95	6.83	2.73

3.3 Scalability

We discuss the comparison of running times of `UNC-1`, `UNC-CE`, and `UNC-IE` methods for naïve Bayes for one iteration of active learning. Given dataset $\mathcal{D} = \left\{ x^{(i)}, y^{(i)} \right\}_1^m$ where, $x^{(i)} \in \mathbb{R}^f$, and $y^{(i)} \in \{+1, -1\}$ is discrete valued. `UNC-1` calculates uncertainty score (measured through Equations 1 or 2). The time complexity of calculating the conditional probabilities $P_\theta(Y|X)$ in each of these equations is
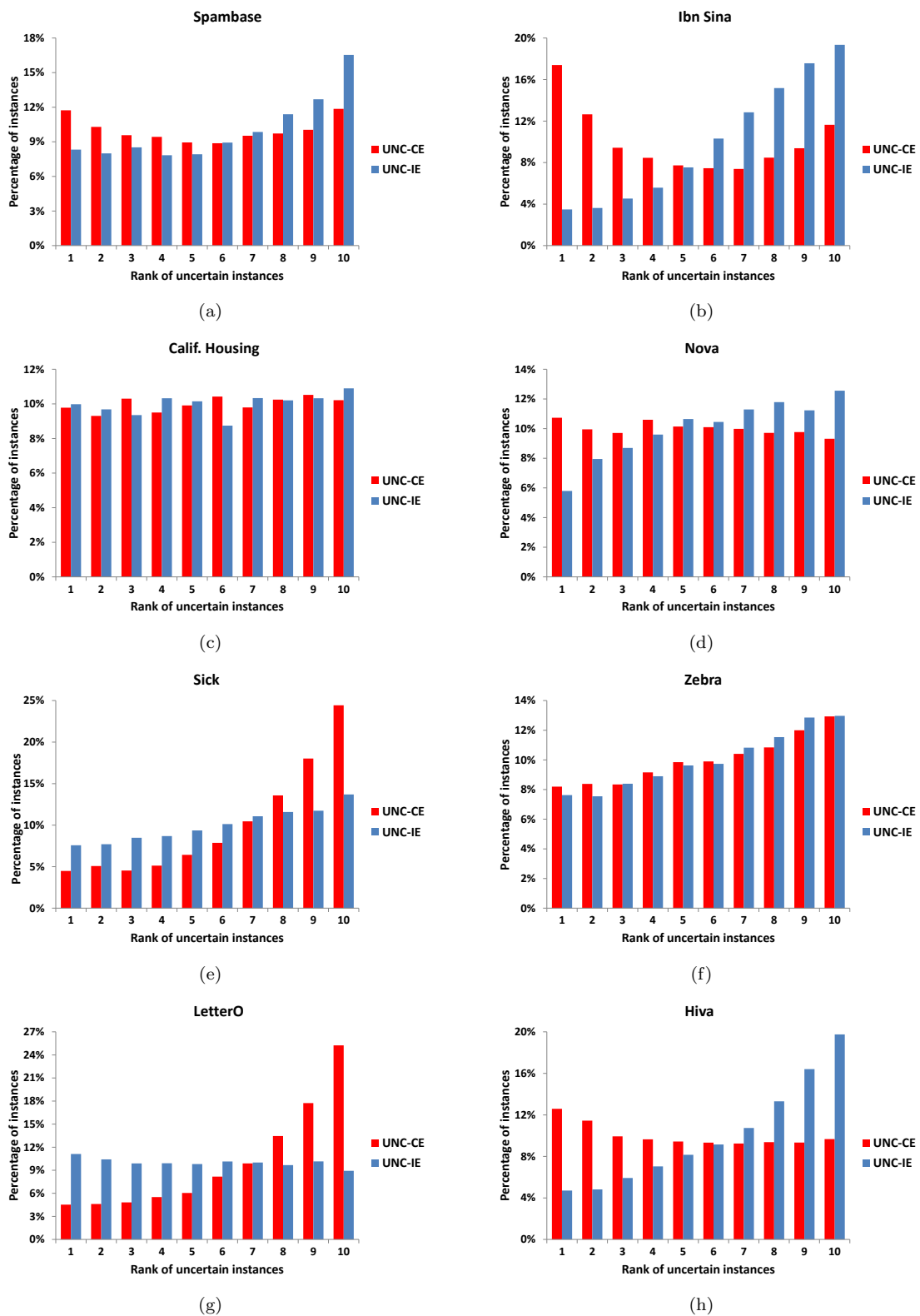
Fig. 4 Histograms showing ranks of uncertain instances selected by UNC-CE and UNC-IE for all eight datasets.

proportional to the number of attributes, which is $O(f)$. Since we compute uncertainty on m subsampled instances, the time complexity of UNC-1 is $O(m \times f)$.

UNC-CE and UNC-IE methods also calculate uncertainty on m instances, which takes time $O(m \times f)$. Additionally, UNC-CE and UNC-IE methods calculate evidence for each attribute of an instance, which again takes time $O(f)$. This additional step is done only for the top t uncertain instances. Hence, the running time of UNC-CE and UNC-IE methods is $O((t + m) \times f)$. Given that t is a small constant ($t << m$), the running times of UNC-CE and UNC-IE are comparable to the running time of UNC-1. Table 5 presents the running times of UNC-1, UNC-CE, and UNC-IE for one iteration of active learning with various t values for three datasets, Nova, Zebra and Hiva. We omit the running times for other five datasets, as the running time per iteration for them is less than 1 second. As presented in Table 1, these three datasets have the highest number of features and thus it is not surprising that the running times are largest for these three datasets. These experiments were run on a Windows 7 machine with Intel Xeon processor (2.4 GHz). The results show that the running times of UNC-CE and UNC-IE are comparable to UNC-1. Moreover, the running times of UNC-CE and UNC-IE do not vary much with different t values. Interestingly, we observe that sometimes UNC-CE and UNC-IE seem to take less time than UNC-1, but these differences are not statistically significant and hence we attribute these differences to variances in the run times due to other uncontrollable factors such as other processes that might be run by the OS. The overall conclusion is that the run time is dominated by the number of features and the additional time cost that UNC-CE and UNC-IE require on top of UNC-1 is negligible.

Table 5 Running times (in seconds) for three datasets for one iteration of active learning, with various t values. We present mean $\pm$ Std. Dev of the running times over 25 trials.

| Dataset	UNC-1	UNC-CE				UNC-IE		
	t=1	t=5	t=10	t=20	t=5	t=10	t=20
Nova	15.02 ± 2.25	14.58 ± 1.44	14.81 ± 1.25	14.95 ± 1.24	15.48 ± 0.98	15.90 ± 0.93	15.72 ± 1.40
Zebra	1.34 ± 0.05	1.33 ± 0.05	1.32 ± 0.05	1.32 ± 0.05	1.36 ± 0.05	1.37 ± 0.06	1.38 ± 0.06
Hiva	1.95 ± 0.05	1.95 ± 0.04	1.95 ± 0.05	1.94 ± 0.05	1.94 ± 0.05	1.95 ± 0.06	1.98 ± 0.06

4 User Study

We designed and ran a user study to investigate whether it is easier or harder for humans to label conflicting-evidence cases versus insufficient-evidence cases. Specifically, we were interested in two measures: i) how long does it take humans to label and ii) how accurate are the humans on their labels for conflicting-evidence cases versus insufficient-evidence cases.

It could very well be that conflicting cases can be harder for humans because they contain conflicting information suggesting both classes, which might confuse humans about the class label. It is also possible for insufficient-evidence cases to be difficult for humans because they do not have enough information, e.g. neutral cases. We note that we define conflicting-evidence and insufficient-evidence uncertainties with respect to the underlying model and not with respect to the expert. Thus, it is possible that the model has conflicting evidence or insufficient evidence but it still might be an easy case for the expert. In this section, we investigate these questions through a user study.

We experimented with IMDB dataset consisting of 50K movie reviews (Maas et al, 2011), as labeling movie reviews does not require much domain expertise and hence it is easier to recruit users for our user study. Moreover, this dataset contains full text of the reviews whereas the other datasets we have

used in Sect. 3 simply consist of feature-value pairs. We trained a multinomial naïve Bayes model, as multinomial naïve Bayes is known to outperform Bernoulli naïve Bayes for text classification (McCallum et al, 1998). The evidences for multinomial naïve Bayes are calculated similar to that of Bernoulli naïve Bayes, which we describe in Sect. 6.

We bootstrapped the multinomial naïve Bayes model with 10 reviews, selecting 5 random reviews from each class and used tf-idf representation of the data. Fig. 5(a) presents the average AUC results of `UNC-CE` and `UNC-IE` strategies over 10 trials simulated using ground truth. Out of the 10 trials, `UNC-CE` wins over `UNC-IE` on 6 of the trials. For the user study, we selected one of the 10 trials, shown in Fig. 5(b), for which `UNC-CE` and `UNC-IE` had the biggest difference in performance because we wanted to test the case where `UNC-CE` and `UNC-IE` had the most difference in impact on the learning. The accuracy of `UNC-CE` after labeling 110 (10 bootstrap + 100 budget) reviews was 73.5% and the accuracy of `UNC-IE` was 67.24%.

(a)   (b)
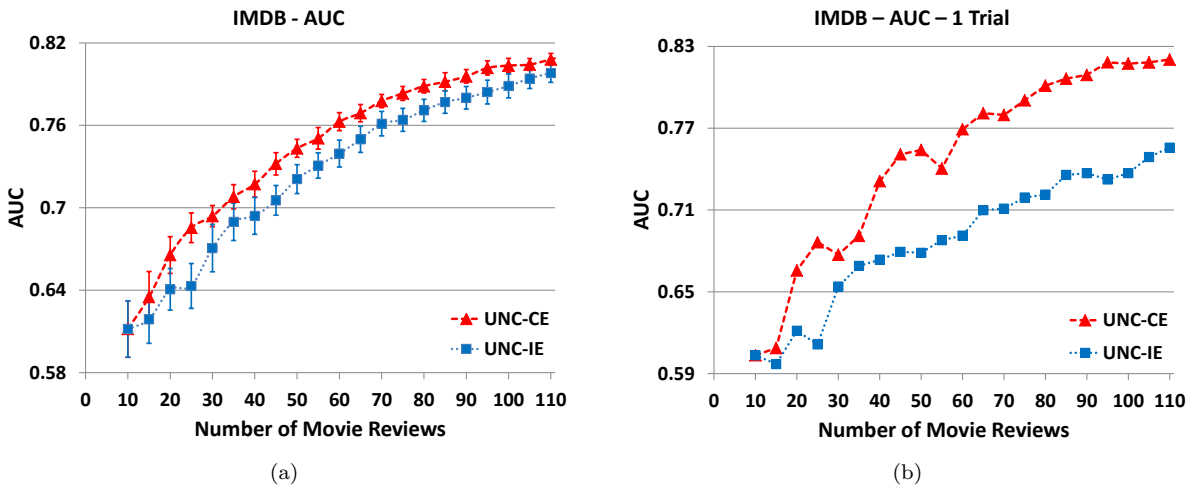
Fig. 5 (a) Average AUC of `UNC-CE` and `UNC-IE` over 10 trials on IMDB dataset. (b) Performance of `UNC-CE` and `UNC-IE` on the trial used in the user study.

We shuffled these 200 movie reviews selected by `UNC-CE` and `UNC-IE` to make sure that the users had no way of determining which was a conflicting versus insufficient evidence case with respect to the underlying model. In fact, users were not told that they were part of a study to distinguish between conflicting versus insufficient evidence cases. They were simply asked to label 200 movie reviews as positive or negative. We had five users for our study and each user was shown movie reviews in the same order. For each movie review, we recorded the response time and annotation (positive/negative). We treated the actual labels as gold standard labels and measured accuracy of the users by comparing their annotations with the gold standard labels.

We first compare whether `UNC-CE` and `UNC-IE` differ on the length of the documents chosen. We observe that the average length of reviews selected by `UNC-CE` was 213.32 and the average length of reviews selected by `UNC-IE` was 205.04. The two-tailed unpaired t-tests between the lengths of `UNC-CE` and `UNC-IE` reviews show that the difference in lengths of `UNC-CE` and `UNC-IE` reviews is not significantly different.

Next, we compare the average time taken by users, in seconds, to label `UNC-CE` and `UNC-IE` reviews in Table 6. We also include the Average User as the mean of all the users in the last row. We observe

that even though users took slightly more time (a few more seconds) on UNC-CE instances than UNC-IE instances, the differences are not statistically significant as measured by two-tailed unpaired t-tests and the p-values are reported in the last column of Table 6.

Table 6 Annotation time of all users on UNC-CE and UNC-IE movie reviews. The t-test results show that annotation times of UNC-CE and UNC-IE reviews are not significantly different. We report the $p-$values obtained using two-tailed unpaired t-tests.

Users	Annotation time of UNC-CE reviews	Annotation time of UNC-IE reviews	$p-$value
User 1	14.27 ± 9.54	13.26 ± 11.03	0.49
User 2	55.40 ± 35.01	52.49 ± 36.64	0.57
User 3	81.03 ± 71.41	74.62 ± 61.55	0.50
User 4	21.86 ± 14.91	20.18 ± 15.82	0.45
User 5	25.79 ± 33.54	26.29 ± 29.68	0.91
Average User	39.57 ± 25.84	37.39 ± 24.25	0.54

Table 7 presents accuracy of the users on the 100 movie reviews selected by UNC-CE and UNC-IE. The accuracy of Average User is the average accuracy of all users. We also present majority vote accuracy which is calculated by taking a majority voting of all users on each movie review. The accuracy of all users, except User 3, was similar for both UNC-CE and UNC-IE reviews.

Table 7 Accuracy of all users on UNC-CE and UNC-IE movie reviews.

Users	Accuracy on UNC-CE reviews	Accuracy on UNC-IE reviews
User 1	95%	93%
User 2	93%	94%
User 3	90%	96%
User 4	95%	95%
User 5	95%	97%
Average User	93.6%	95%
Majority vote accuracy	96%	96%

We plot the same Fig. 5(b) again, this time the x-axis is not the number of instances but rather the average time it took the 5 users (i.e., the Average User's time). Fig. 6(a) shows the results using the ground-truth labels and Fig. 6(b) shows the results using majority vote labels. [3] This result shows that even though labeling UNC-CE reviews takes slightly more time than UNC-IE reviews, it is still worth labeling reviews using UNC-CE strategy.

Overall, our user study focused on sentiment classification. Though we cannot claim that our results carry over to other document classification tasks or other domains, we conclude that we did not observe significant differences between UNC-CE and UNC-IE instances in terms of the annotation time and labeling difficulty for the sentiment classification task.

---

[3] This figure does not correspond to a real-time simulation of active learning with users. When the user-provided labels are used, the underlying active learning strategy, whether it be UNC-CE or UNC-IE, would potentially take a different path per user based on their labels. Then, each user would potentially differ on the documents they label, and therefore meaningful comparisons of time and accuracy across users would not be possible.
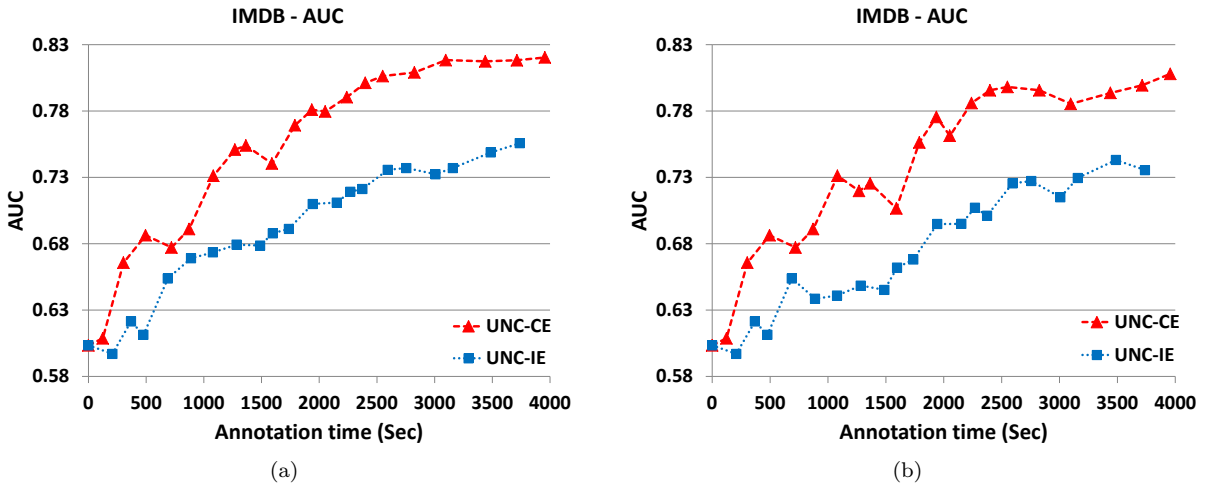
Fig. 6 (a) Performance comparison of `UNC-CE` and `UNC-IE` strategies based on the annotation time of Average User using ground-truth labels. (b) Performance comparison of `UNC-CE` and `UNC-IE` strategies based on the annotation time of Average User and using majority vote labels.

5 Analytical and Empirical Justifications

Extensive experiments with real-world datasets presented in Sect. 3.2 clearly show that `UNC-CE` provides significant improvements over `UNC-IE`. This is a bit startling because one would expect that the model would benefit from both the `UNC-CE` cases and `UNC-IE` cases. When the conflicting, `UNC-CE`, cases are annotated, the model would have a chance to correct its perceived conflict, and when the inconclusive, `UNC-IE`, cases are annotated, the model would learn about new feature-value class correlations that it did not know before. In this section, we provide both analytical and empirical results that shed light on why `UNC-CE` often outperforms `UNC-IE`. Specifically,

– We show both analytically and empirically that `UNC-CE` cases have lower density, with respect to the model trained on the labeled data, than the `UNC-IE` cases. Density of an instance, $x^{(i)}$, is defined as the probability distribution, $P(x^{(i)})$, with respect to the model trained on the current training data.
– We show empirically that the model has higher variance on the `UNC-CE` cases than on the `UNC-IE` cases.

These two results suggest that the conflict perceived by the model is supported by less amount of training data than the insufficiency of the evidences. Put another way, there is less labeled data that supports the conflict and there is more labeled data that supports the inconclusiveness. This is further supported by the finding that `UNC-CE` cases have higher variance than `UNC-IE` cases. That is, the parameter values that support conflict have higher variance because they rely on smaller amount of labeled data. Therefore, the model is more likely to be incorrect in its decision that the evidence is conflicting than its decision that the evidence is inconclusive.

This is not to say that the `UNC-IE` cases are totally useless. Even though `UNC-IE` cases are supported by more labeled data than the `UNC-CE` cases, the total amount of labeled data is still fairly small in active learning settings. Therefore, the model is likely to be incorrect in its decision that the case is inconclusive. However, the `UNC-CE` cases have even less support than the `UNC-IE` cases and thus the model is often better off labeling more of the `UNC-CE` cases.

5.1 Analytical Justification

For simplicity, we first prove the density argument for binary variables using a two-attributes case where out of four possible cases, one is `UNC-CE` and the other is `UNC-IE`. We then provide explanation of density argument for continuous attributes.

5.1.1 Binary Attributes

Assume we have a single attribute, X_1, that is binary with $\langle T, F \rangle$. Similarly, the class variable Y is binary with $\langle -1, +1 \rangle$. In this section, we prove that i) $X_1 = T$ and $X_1 = F$ cannot provide evidence for the same class at the same time, ii) if $X_1 = T$ provides evidence for one class, then $X_1 = F$ has to provide evidence for the opposing class, and finally iii) the amount of evidence that $X_1 = T$ provides for one class can be larger/smaller than the evidence $X_1 = F$ provides for the opposing class. These three properties will be needed to prove the density argument for the two-attributes case. Let

$$P(X_1 = T | Y = +1) = p; P(X_1 = F | Y = +1) = 1 - p$$
$$P(X_1 = T | Y = -1) = q; P(X_1 = F | Y = -1) = 1 - q$$

The following propositions hold when both X_1 and Y are binary.

Proposition 1: If $X_1 = T$ provides evidence for $Y = +1$, then $X_1 = F$ cannot provide evidence for $Y = +1$ at the same time.

Proof. Without loss of generality, assume $p > q$. Then, $X_1 = T$ provides evidence for $Y = +1$ and the magnitude of the evidence is $\frac{p}{q}$. Can $X_1 = F$ provide evidence for $Y = +1$ at the same time? That is, when $p > q$, can $\frac{1-p}{1-q}$ be greater than 1? The answer is obviously no and hence two different values of X_1 cannot provide evidence for the same class at the same time.

Proposition 2: If $X_1 = T$ provides evidence for one class then $X_1 = F$ *has* to provide evidence for the other class.

Proof. When $X_1 = T$ provides evidence for one class, is it possible that $X_1 = F$ provides evidence for no class? That is, is it possible to have $\frac{p}{q} \neq 1$ and $\frac{1-p}{1-q} = 1$? This is obviously impossible, and hence if $X_1 = T$ provides evidence for one class then $X_1 = F$ *has* to provide evidence for some class. Given proposition 1, we know that $X_1 = F$ cannot provide evidence for the class that $X_1 = T$ supports. Therefore, if $X_1 = T$ supports one class, then $X_1 = F$ has to support the other class.

Proposition 3: One value of an attribute can provide a greater evidence for one class than the evidence the other value of the same attribute provides for the other class.

Proof. Without loss of generality, assume $\frac{p}{q} > 1$. Then, $X_1 = T$ provides evidence for $Y = +1$. Hence $\frac{1-q}{1-p} > 1$ and $X_1 = F$ provides evidence for $Y = -1$. The evidence that $X_1 = T$ provides for $Y = +1$ is greater than the evidence $X_1 = F$ provides for $Y = -1$, that is, $\frac{p}{q} > \frac{1-q}{1-p}$, if and only if $p = q + \epsilon \leq 0.5$ for $\epsilon > 0$ or $p = 0.5 + \alpha$ and $q = 0.5 - \beta$ for $0 < \alpha < \beta < 0.5$.

For the two-attributes case, assume we have two binary attributes, X_1 and X_2. In this case, there are four possible instances (e.g., $\langle X_1 = T, X_2 = T \rangle$, $\langle X_1 = T, X_2 = F \rangle$, etc.). To compare `UNC-CE` and `UNC-IE` methods, we need the model to be uncertain on at least two of these instances and we want one of them to be a conflicting-evidence case and the other one to be an insufficient-evidence case. Assume the following distributions for a naïve Bayes classifier:

$$P(X_1 = T|Y = +1) = p; P(X_1 = F|Y = +1) = 1 - p$$
$$P(X_1 = T|Y = -1) = q; P(X_1 = F|Y = -1) = 1 - q$$
$$P(X_2 = T|Y = +1) = r; P(X_2 = F|Y = +1) = 1 - r$$
$$P(X_2 = T|Y = -1) = s; P(X_2 = F|Y = -1) = 1 - s$$

Assume that the uncertain instances are $\langle X_1 = T, X_2 = T \rangle$ and $\langle X_1 = F, X_2 = F \rangle$. That is:

$$\frac{P(Y = +1)P(X_1 = T|Y = +1)P(X_2 = T|Y = +1)}{P(Y = -1)P(X_1 = T|Y = -1)P(X_2 = T|Y = -1)} \approx 1$$
$$\frac{P(Y = +1)P(X_1 = F|Y = +1)P(X_2 = F|Y = +1)}{P(Y = -1)P(X_1 = F|Y = -1)P(X_2 = F|Y = -1)} \approx 1$$

Without loss of generality, assume $X_1 = T$ provides evidence for $Y = +1$. Then, Propositions 1 and 2 above show that $X_1 = F$ provides evidence for $Y = -1$. Assuming $P(Y)$ is uniform with 0.5, for the instance $\langle X_1 = T, X_2 = T \rangle$ to be uncertain, $X_2 = T$ must provide evidence for $Y = -1$ and this evidence must be roughly equal to the evidence that $X_1 = T$ provides for $Y = +1$. Invoking propositions 1 and 2 again, $X_2 = F$ then must provide evidence for $Y = +1$ and for $\langle X_1 = F, X_2 = F \rangle$ to be uncertain, the evidence $X_1 = F$ provides for $Y = -1$ must be roughly equal to the evidence $X_2 = F$ provides for $Y = +1$.

Without loss of generality, assume $\langle X_1 = T, X_2 = T \rangle$ is the UNC-CE instance and $\langle X_1 = F, X_2 = F \rangle$ is the UNC-IE instance. Then, for both instances to be uncertain, and for $\langle X_1 = T, X_2 = T \rangle$ to be the conflicting case as opposed to $\langle X_1 = F, X_2 = F \rangle$, we need

$$\frac{p}{q} \approx \frac{s}{r} > \frac{1 - q}{1 - p} \approx \frac{1 - r}{1 - s}$$

Proposition 4: The density of UNC-CE instance *with respect to* the naïve Bayes model is less than the density of UNC-IE instance, i.e. $P(X_1 = T, X_2 = T) < P(X_1 = F, X_2 = F)$.

Proof. Assume that $P(Y)$ is uniform, $P(Y = +1) = P(Y = -1) = 0.5$. We need to prove that

$$0.5 \times p \times r + 0.5 \times q \times s < 0.5 \times (1 - p) \times (1 - r) + 0.5 \times (1 - q) \times (1 - s)$$

$$0.5 \times p \times r + 0.5 \times q \times s \overset{?}{<} 0.5 \times (1 - p) \times (1 - r) + 0.5 \times (1 - q) \times (1 - s)$$
$$p \times r + q \times s \overset{?}{<} (1 - p) \times (1 - r) + (1 - q) \times (1 - s)$$
$$p \times r + q \times s \overset{?}{<} 1 - r - p + p \times r + 1 - s - q + q \times s$$
$$0 \overset{?}{<} 2 - r - p - s - q$$
$$r + p + s + q \overset{?}{<} 2$$

Because $\frac{p}{q} > \frac{1-q}{1-p}$ and as we have shown in Proposition 3, either $p = q + \epsilon \leq 0.5$ for $\epsilon > 0$ or $p = 0.5 + \alpha$ and $q = 0.5 - \beta$ for $0 < \alpha < \beta < 0.5$. Similar arguments apply to s and r: either $s = r + \epsilon \leq 0.5$ for $\epsilon > 0$ or $s = 0.5 + \alpha$ and $r = 0.5 - \beta$ for $0 < \alpha < \beta < 0.5$.

Case 1: $p = q + \epsilon \leq 0.5$ for $\epsilon > 0$. Then $p + q < 1$. Similarly, if $s = r + \epsilon \leq 0.5$ for $\epsilon > 0$, then $s + r < 1$.

Case 2: $p = 0.5 + \alpha$ and $q = 0.5 - \beta$ for $0 < \alpha < \beta < 0.5$. Then $p + q = 0.5 + \alpha + 0.5 - \beta = 1 + \alpha - \beta < 1$. Similarly, $s + r = 0.5 + \alpha + 0.5 - \beta = 1 + \alpha - \beta < 1$.

Since in both cases, $p + q < 1$ and $s + r < 1$, we conclude that $p + q + r + s < 2$, proving that the density with respect to the underlying naïve Bayes model is lower for the UNC-CE case than the UNC-IE case. Our proof assumed that $P(Y)$ was uniform; the proposition holds when $P(Y)$ is not uniform and the proof is similar. Moreover, for simplicity, our proof focused on the two-attributes case. The same arguments can be extended to multiple-attributes case by induction.

5.1.2 Continuous Attributes

In this section we investigate the density hypothesis for continuous attributes. For continuous attributes, Gaussian naïve Bayes assumes that within each class, the continuous attributes are normally distributed:

$$p(x|Y) = \mathcal{N}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

For simplicity of exposition, consider a training data with two continuous attributes, X_1 and X_2, and a binary class variable, Y with $\langle -1, +1 \rangle$. Let the mean of attribute X_1 for class $+1$ be $\mu_{1,+1}$ and mean of attribute X_1 for class -1 be $\mu_{1,-1}$. Similarly, let mean of attribute X_2 for class $+1$ be $\mu_{2,+1}$ and mean of attribute X_2 for class -1 be $\mu_{2,-1}$. Let the standard deviation of attribute X_1 for class $+1$ be $\sigma_{1,+1}$ and standard deviation of attribute X_1 for class -1 be $\sigma_{1,-1}$. Similarly, let standard deviation of attribute X_2 for class $+1$ be $\sigma_{2,+1}$ and standard deviation of attribute X_2 for class -1 be $\sigma_{2,-1}$. For each class and attribute, Gaussian naïve Bayes estimates the conditional probability of attribute given class as:

$$p(X_1|Y = +1) = \mathcal{N}(\mu_{1,+1}, \sigma_{1,+1})$$
$$p(X_1|Y = -1) = \mathcal{N}(\mu_{1,-1}, \sigma_{1,-1})$$
$$p(X_2|Y = +1) = \mathcal{N}(\mu_{2,+1}, \sigma_{2,+1})$$
$$p(X_2|Y = -1) = \mathcal{N}(\mu_{2,-1}, \sigma_{2,-1})$$

Assume $\mu_{1,+1} = \mu_{2,+1} = a$ and $\mu_{1,-1} = \mu_{2,-1} = b$, where $b > a$. This can be easily achieved by rotating and shifting the axes. For simplicity, assume that both attributes have equal variance in both classes, i.e. $\sigma_{1,+1} = \sigma_{1,-1} = \sigma_{2,+1} = \sigma_{2,-1} = \sigma$ (the case where each class and feature value pair has unequal variances is similar). Hence the data for class $+1$ is centered around the point $\langle a, a \rangle$ and the data for class -1 is centered around the point $\langle b, b \rangle$. Fig. 7 illustrates these points for the two classes. The decision boundary represents the line where an instance has equal probability, 0.5, of belonging to each class. We consider two instances on the decision boundary, where one instance is $\langle \frac{a+b}{2}, \frac{a+b}{2} \rangle$ and the other is $\langle c, d \rangle$, assuming $c < \frac{a+b}{2}$ and $d > \frac{a+b}{2}$.

Next we provide analytical justification showing that conflicting cases have higher evidence but lower density in the training data, whereas insufficient-evidence cases have lower evidence and higher density in the training data.

Proposition 5: Instance $\langle c, d \rangle$ has higher total evidence than instance $\langle \frac{a+b}{2}, \frac{a+b}{2} \rangle$.

Proof. First, we show how the evidences for $+1$ (or -1) class can be computed. The evidence provided by attribute X_f for class $+1$ using Gaussian naïve Bayes is computed as:
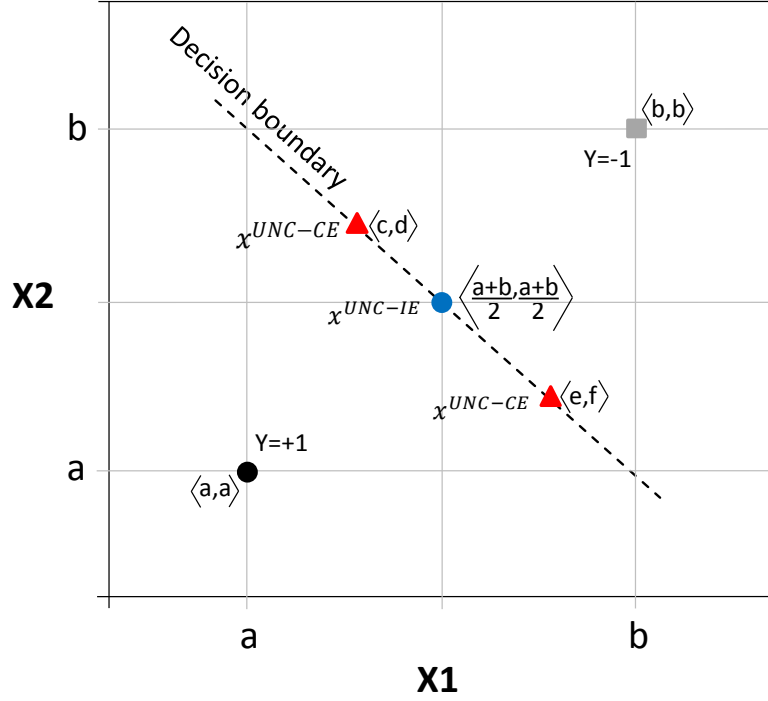
Fig. 7 Analysis of Gaussian naïve Bayes using two continuous attributes, X_1 and X_2. The mean of both attributes for class $+1$ is a, and the mean of both attributes for class -1 is b. We consider two instances, $\langle \frac{a+b}{2}, \frac{a+b}{2} \rangle$ and $\langle c, d \rangle$ on the decision boundary and we prove that $\langle \frac{a+b}{2}, \frac{a+b}{2} \rangle$ is insufficient-evidence uncertain instance and we refer to it as x^{UNC-IE} on this graph, and $\langle c, d \rangle$ is conflicting-evidence uncertain instance and we refer to it as x^{UNC-CE} on this graph.

$$\frac{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_f - \mu_{f,+1})^2}{2\sigma^2}}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_f - \mu_{f,-1})^2}{2\sigma^2}}} = e^{\frac{-(X_f - \mu_{f,+1})^2 + (X_f - \mu_{f,-1})^2}{2\sigma^2}}$$

$$= e^{\frac{(X_f - \mu_{f,-1} + X_f - \mu_{f,+1})(X_f - \mu_{f,-1} - X_f + \mu_{f,+1})}{2\sigma^2}}$$

$$= e^{\frac{(2X_f - \mu_{f,-1} - \mu_{f,+1})(\mu_{f,+1} - \mu_{f,-1})}{2\sigma^2}}$$

For class -1, this ratio is reversed, hence the evidence provided by attribute X_f for the class -1 is:

$$e^{\frac{(2X_f - \mu_{f,-1} - \mu_{f,+1})(\mu_{f,-1} - \mu_{f,+1})}{2\sigma^2}}$$

First, we compute the evidences for instance $\langle \frac{a+b}{2}, \frac{a+b}{2} \rangle$. The evidence that attribute X_1 of instance $\langle \frac{a+b}{2}, \frac{a+b}{2} \rangle$ provides for class $+1$ is:

$$e^{\frac{(2(\frac{a+b}{2}) - a - b)(a - b)}{2\sigma^2}} = e^0 = 1$$

That is, $X_1 = \frac{a+b}{2}$ does not provide evidence for either class, because $\frac{P(X_1 = \frac{a+b}{2}|+1)}{P(X_1 = \frac{a+b}{2}|-1)} = 1$. The same argument applies to $X_2 = \frac{a+b}{2}$. The overall evidence provided by attributes of instance $\langle \frac{a+b}{2}, \frac{a+b}{2} \rangle$ using Eq. 9 is $1 \times 1 = 1$.

Next we compute the evidences for instance $\langle c, d \rangle$. Note that c is closer to class $+1$ and d is closer to class -1. The evidence that $X_1 = c$ provides for class $+1$ is:

$$e^{\frac{(2c-a-b)(a-b)}{2\sigma^2}}$$

Since $c < \frac{a+b}{2}$ and $a < b$, this evidence is greater than 1. The evidence that $X_2 = d$ provides for class -1 is:

$$e^{\frac{(2d-a-b)(b-a)}{2\sigma^2}}$$

Since $d > \frac{a+b}{2}$ and $b > a$, this evidence is greater than 1. The total evidence provided by attributes of instance $\langle c, d \rangle$ using Eq. 9 is:

$$e^{\frac{(2c-a-b)(a-b)}{2\sigma^2}} \times e^{\frac{(2d-a-b)(b-a)}{2\sigma^2}}$$

which is greater than 1, whereas the total evidence provided by attributes of $\langle \frac{a+b}{2}, \frac{a+b}{2} \rangle$ is equal to 1.

Similar reasoning applies to the instance $\langle e, f \rangle$ in Fig. 7. We conclude that as we move on the decision boundary away from its center, i.e. move away from $\langle \frac{a+b}{2}, \frac{a+b}{2} \rangle$ in the direction of $\langle c, d \rangle$ (or $\langle e, f \rangle$), the evidences for each class get higher and hence the conflict grows.

Before we prove the density argument that conflicting-evidence cases have lower density compared to the insufficient-evidence cases, we first establish a relationship among c, d, a, and b. Note that for instance $\langle c, d \rangle$ to be uncertain, the evidence for class $+1$ must be equal to the evidence for class -1. Hence,

$$e^{\frac{(2c-a-b)(a-b)}{2\sigma^2}} = e^{\frac{(2d-a-b)(b-a)}{2\sigma^2}}$$
$$\therefore 2c - a - b = a + b - 2d$$
$$c + d = a + b$$

Proposition 6: The density of instance $\langle c, d \rangle$ with respect to the underlying model is lower than the density of instance $\langle \frac{a+b}{2}, \frac{a+b}{2} \rangle$.

Proof. Density of instance $\langle X_1, X_2 \rangle$ with respect to the underlying model can be computed as follows:

$$P(X_1, X_2) = P(X_1, X_2, +1) + P(X_1, X_2, -1)$$
$$= P(+1)P(X_1, X_2 | +1) + P(-1)P(X_1, X_2 | -1)$$

Naïve Bayes assumes that attributes are conditionally independent given class, hence,

$$P(X_1, X_2) = P(+1)P(X_1|+1)P(X_2|+1) + P(-1)P(X_1|-1)(X_2|-1)$$

$$= P(+1) \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_1-\mu_{1,+1})^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_2-\mu_{2,+1})^2}{2\sigma^2}} +$$

$$P(-1) \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_1-\mu_{1,-1})^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_2-\mu_{2,-1})^2}{2\sigma^2}}$$

Assuming $P(+1) = P(-1) = 0.5$,

$$P(X_1, X_2) = \frac{1}{2} \times \frac{1}{\sqrt{2\pi\sigma^2}} \left(e^{-\frac{(X_1-\mu_{1,+1})^2+(X_2-\mu_{1,+1})^2}{2\sigma^2}} + e^{-\frac{(X_1-\mu_{1,-1})^2+(X_2-\mu_{1,-1})^2}{2\sigma^2}} \right)$$

Density of instance $\langle \frac{a+b}{2}, \frac{a+b}{2} \rangle$ is:

$$\frac{1}{2} \times \frac{1}{\sqrt{2\pi\sigma^2}} \left(e^{-\frac{(\frac{a+b}{2}-a)^2+(\frac{a+b}{2}-a)^2}{2\sigma^2}} + e^{-\frac{(\frac{a+b}{2}-b)^2+(\frac{a+b}{2}-b)^2}{2\sigma^2}} \right)$$

$$= \frac{1}{2} \times \frac{1}{\sqrt{2\pi\sigma^2}} \left(e^{-\frac{(\frac{b-a}{2})^2}{2\sigma^2}} + e^{-\frac{(\frac{a-b}{2})^2}{2\sigma^2}} \right)$$

$$= \frac{1}{2} \times \frac{1}{\sqrt{2\pi\sigma^2}} \times 2e^{-\frac{(\frac{b-a}{2})^2}{2\sigma^2}}$$

Density of instance $\langle c, d \rangle$ is:

$$\frac{1}{2} \times \frac{1}{\sqrt{2\pi\sigma^2}} \left(e^{-\frac{(c-a)^2+(d-a)^2}{2\sigma^2}} + e^{-\frac{(c-b)^2+(d-b)^2}{2\sigma^2}} \right)$$

First, note that $(c-a)^2 + (d-a)^2 = (c-b)^2 + (d-b)^2$.

$$(c-a)^2 - (c-b)^2 \overset{?}{=} (d-b)^2 - (d-a)^2$$

$$(c-a+c-b)(c-a-c+b) \overset{?}{=} (d-b+d-a)(d-b-d+a)$$

$$(2c-a-b)(-a+b) \overset{?}{=} (2d-b-a)(-b+a)$$

$$2c-a-b \overset{?}{=} b+a-2d$$

$$c+d \overset{?}{=} a+b$$

We earlier established relationship among c, d, a, and b and proved that $c+d = a+b$. Therefore, the density of instance $\langle c, d \rangle$ is:

$$\frac{1}{2} \times \frac{1}{\sqrt{2\pi\sigma^2}} 2e^{-\frac{(c-a)^2+(d-a)^2}{2\sigma^2}}$$

Next, we test whether density of instance $\langle \frac{a+b}{2}, \frac{a+b}{2} \rangle$ is higher than density of instance $\langle c, d \rangle$.

$$\frac{1}{2} \times \frac{1}{\sqrt{2\pi\sigma^2}} \times 2e^{-\frac{(\frac{b-a}{2})^2}{2\sigma^2}} \overset{?}{>} \frac{1}{2} \times \frac{1}{\sqrt{2\pi\sigma^2}} 2e^{-\frac{(c-a)^2+(d-a)^2}{2\sigma^2}}$$

$$\frac{(b-a)^2}{2} \overset{?}{<} (c-a)^2 + (d-a)^2$$

$$(b-a)^2 \overset{?}{<} 2(c-a)^2 + 2(d-a)^2$$

Since $c + d = a + b$, assume $c = a + \epsilon$ and $d = b - \epsilon$, where ϵ is any real number.

$$(b-a)^2 \overset{?}{<} 2(a+\epsilon-a)^2 + 2(b-\epsilon-a)^2$$

$$(b-a)^2 \overset{?}{<} 2\epsilon^2 + 2(b-a)^2 + 2\epsilon^2 - 4(b-a\epsilon)$$

$$0 \overset{?}{<} 4\epsilon^2 + (b-a)^2 + -4(b-a\epsilon)$$

$$0 \overset{?}{<} (2\epsilon - b + a)^2$$

For any real numbers, a, b, and c, $(2\epsilon - b + a)^2$ will always be greater than 0, except when $\epsilon = \frac{b+a}{2}$, $(2\epsilon - b + a)^2$ will be equal to 0. When $\epsilon = \frac{b+a}{2}$, $c = a + \frac{b-a}{2}$, i.e. $c = \frac{a+b}{2}$. For any other value of ϵ, instance $\langle \frac{a+b}{2}, \frac{a+b}{2} \rangle$ has a higher density, with respect to the underlying model, than instance $\langle c, d \rangle$.

5.2 Empirical Justifications

We have shown that the UNC-CE case has lower density than the UNC-IE case, with respect to the underlying naïve Bayes model. Our proof assumed that the instances were nearly perfectly uncertain, i.e. $P(X|Y = +1) = P(X|Y = -1) = 0.5$. In reality, however, it is impractical to assume that the instances lie perfectly on the decision boundary. To analyze such cases, we provide an empirical study to investigate the correlation between density and evidence for instances that are close to decision boundary but not necessarily on the decision boundary of the model.

We created a synthetic dataset using a Bernoulli Naïve Bayes model where the number of features was 10. We assumed that each parameter had a Beta prior, and hence the posterior was also a Beta distribution. Note that even though the joint posterior distribution $P(Y, X|\mathcal{L})$ has a closed-form solution, computing the conditional $P(Y|X, \mathcal{L})$ requires us to resort to sampling. Therefore, rather than plugging in the mean of the posterior distributions for $P(Y|\mathcal{L})$ and $P(X|Y, \mathcal{L})$, we instead sampled their values from their posterior distributions, which gave us a sample over $P(Y|X, \mathcal{L})$, rather than a single point estimate. Using this sample, we computed the variance of $P(Y|X, \mathcal{L})$.

We tested if, how, and how much the evidence, density, and variance are correlated for the top uncertain instances. We used Eq. 2 to compute the uncertainty score of all instances $x^{(i)} \in \mathcal{U}$ and considered instances above the threshold of 0.45 uncertainty score to be the uncertain instances. We computed the evidence, which we earlier defined as $E_{+1}(x^{(i)}) \times E_{-1}(x^{(i)})$, for each uncertain instance $x^{(i)}$, and ranked them in increasing order of evidence. Let this ranking be r_e. We compared this ranking with the ranking with respect to variance, r_v, and with the ranking with respect to density, r_d.

We computed the Spearman rank correlation between the evidence-based ranking, r_e, and the variance-based ranking, r_v. We also computed the Spearman rank correlation between the evidence-based ranking, r_e, and the density-based ranking, r_d. We computed the correlations for various sizes of

Table 8 Spearman rank correlations between evidence and density, and evidence and prediction variance, with respect to the model trained on $\mathcal{L}$.

$|\mathcal{L}|$	Evidence's Correlation with			
	Density		Variance	
	Mean	Std. Dev	Mean	Std. Dev
20	-0.84	0.0353	0.91	0.0139
40	-0.93	0.0031	0.96	0.0018
60	-0.94	0.0025	0.97	0.0006
80	-0.95	0.0025	0.97	0.0006
100	-0.92	0.0092	0.95	0.0093

labeled data, $\mathcal{L}$. We repeated each experiment 10 times, each time randomly choosing the labeled data $\mathcal{L}$. We report the mean and standard deviation of the correlations over the 10 trials.

Table 8 presents the results for Spearman rank correlations between evidence and density, and between evidence and variance of the posterior predictive distribution, of the uncertain instances for various training data sizes, $|\mathcal{L}|$. These results clearly show that the amount of evidence the model has on uncertain instances and the densities of these uncertain instances with respect to the model are highly negatively correlated (ranging between -0.84 and -0.95), providing empirical evidence that uncertain instances with higher evidence (UNC-CE instances) have lower density in the training data than the uncertain instances with lower evidence (UNC-IE instances). These results further show that the Spearman rank correlation between r_e and r_v is positive and quite high, ranging from 0.91 to 0.97 for various training data sizes, showing that UNC-CE cases have higher variance than the UNC-IE cases.

In Fig. 8 we plot the histograms of the posterior predictive distributions $P(Y = +1|X, \mathcal{L})$ for two instances for which the model is uncertain for different reasons: conflicting vs. inconclusive evidences. In both cases, the model is equally uncertain on X where the mean of $P(Y = +1|X, \mathcal{L})$ is 0.49. However, UNC-CE instance (the red histogram) has twice the variance of the UNC-IE instance (the blue histogram), 0.10 versus 0.05 respectively. Regular uncertainty sampling for active learning would not make a distinction between these two instances as both have equally high uncertainty of 0.49, but UNC-CE strategy would prefer the high variance one and the UNC-IE strategy would prefer the low variance one.
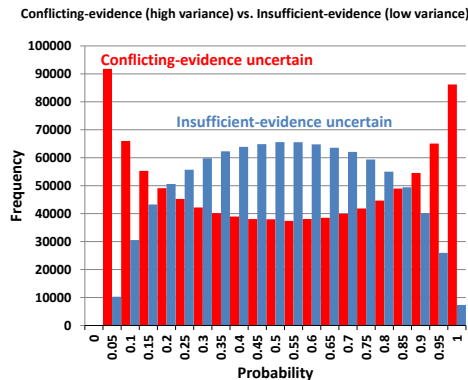
Fig. 8 The histogram of $P(Y = +1|X, \mathcal{L})$ for two instances that are uncertain for two different reasons: conflicting-evidence vs. insufficient-evidence.

We have seen that the underlying model has higher variance on `UNC-CE` cases. Next, we compare `UNC-CE` and `UNC-IE` strategies to query-by-committee strategy (Seung et al, 1992), which chooses instances on which the model has the highest prediction variance.

5.3 Comparison to query-by-committee

Query-by-committee (`QBC`) (Seung et al, 1992) is another frequently used baseline in active learning. `QBC` selects instances that reduce the version space size of the underlying model class (Mitchell, 1982). A committee of classifiers is formed by sampling hypotheses from the version space, but since this is not always possible, an approximate version of `QBC` can be formed by technique known as bagging which is described in (Abe and Mamitsuka, 1998) and selects instances on which the committee disagrees the most. The two most common approaches to measure the disagreement between committee members are margin of disagreement, i.e. the difference between number of votes for the most popular label and number of votes for the next most popular label (Melville and Mooney, 2004), and vote entropy (Dagan and Engelson, 1995). Vote entropy is defined as:

$$x^* = \arg\max_{x^{(i)} \in \mathcal{U}} - \sum_{y \in Y} \frac{V(y)}{C} log \frac{V(y)}{C} \tag{15}$$

where y ranges over all possible labels in Y, $V(y)$ is the number of votes that a label receives from the committee members, and C is the committee size.

We built a committee of 10 classifiers using bagging technique described in (Abe and Mamitsuka, 1998) and used vote entropy (Dagan and Engelson, 1995) as a measure of informativeness of instances. Figs. 9 and 10 present the learning curves comparing `UNC-CE` and `UNC-IE` with $t = 10$ to `QBC`. These results show that for most datasets and measures, `UNC-CE` outperforms `QBC` whereas `UNC-IE` is worse than `QBC`. Table 9 presents the t-test results comparing `UNC-1`, `UNC-CE`, and `UNC-IE` to `QBC`. For AUC measure, `UNC-CE` significantly wins over `QBC` on seven datasets and loses on one (Spambase), whereas `UNC-IE` loses to `QBC` on all datasets except Hiva. For accuracy, `UNC-CE` significantly outperforms `QBC` on three datasets and loses on one (Ibn Sina), and for F1, it wins on three datasets and loses on one (LetterO). `UNC-IE` loses to `QBC` for both accuracy and F1 measures for all datasets.

Table 9 `UNC-1`, `UNC-CE`, and `UNC-IE` versus `QBC`. Number of datasets on which `UNC-1`, `UNC-CE`, and `UNC-IE` significantly Win (W), Tie (T), or Lose (L) compared to `QBC` baseline.

QBC baseline	AUC	ACCU	F1
Method	W/T/L	W/T/L	W/T/L
UNC-1	3/0/**5**	**2**/0/**2**	**4**/0/0
UNC-CE	**7**/0/1	**3**/0/1	**3**/0/1
UNC-IE	1/0/**7**	0/0/**4**	0/0/**4**

5.4 Discussion

We presented both analytical and empirical results showing that the conflicting cases have lower density, with respect to the underlying model, than the inconclusive cases. That is, the perceived conflict is
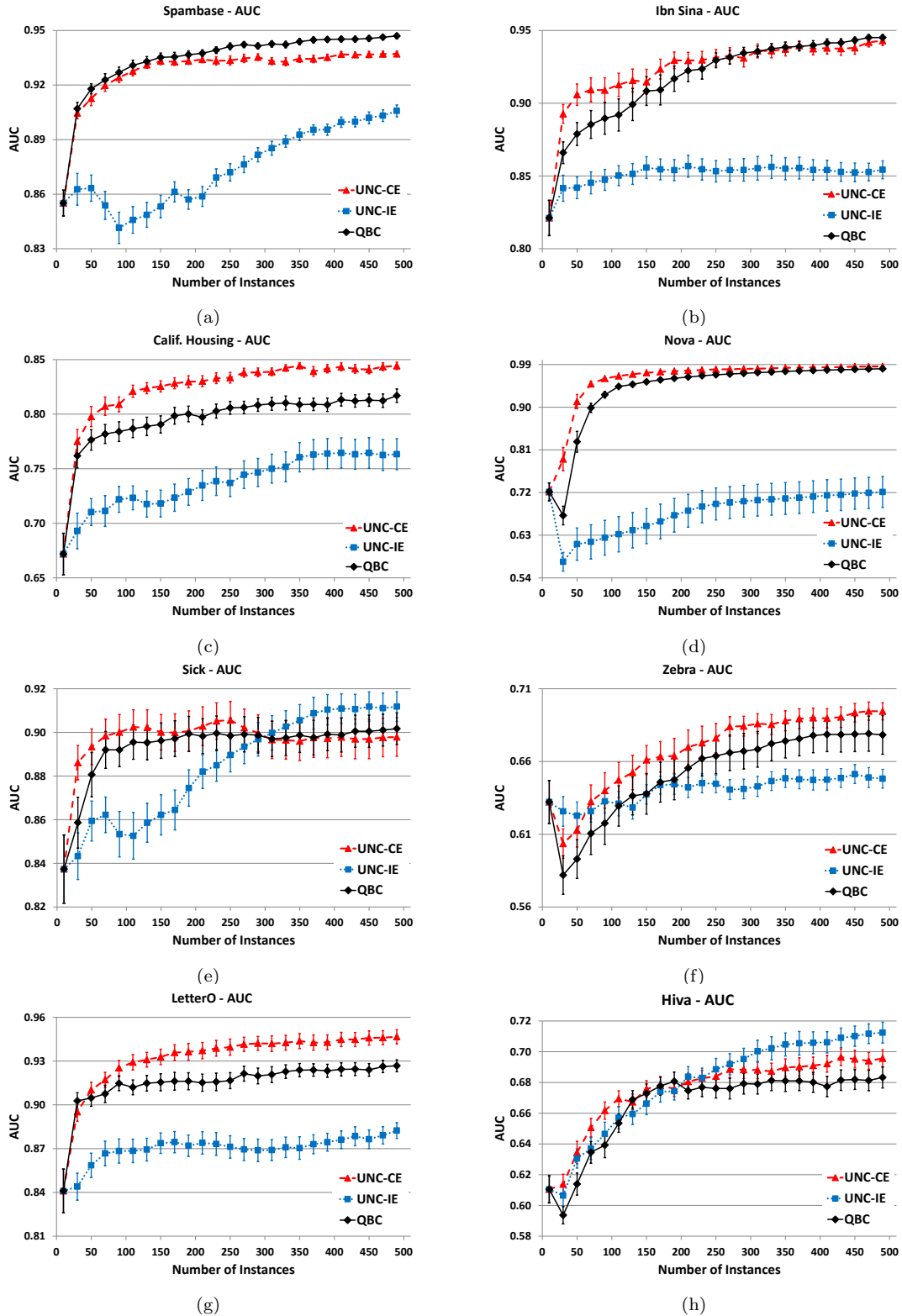
Fig. 9 AUC results for all eight datasets. UNC-CE outperforms QBC on seven out of eight datasets ((b), (c), (d), (e), (f), (g), and (h)) and loses on Spambase dataset (a). UNC-IE loses to QBC on seven out of eight datasets ((a), (b), (c), (d), (e), (f), and (g), and wins on Hiva dataset (h).
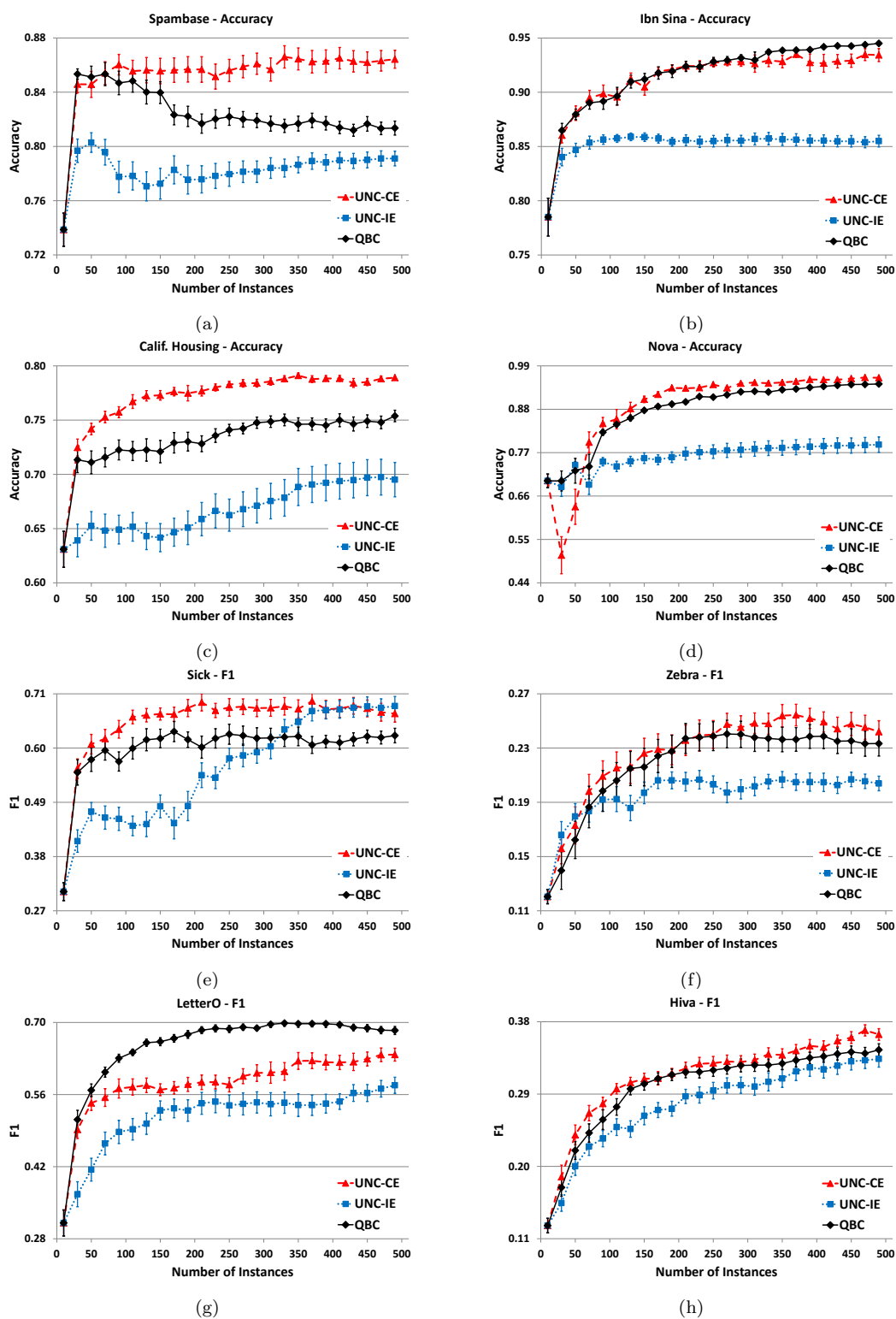
Fig. 10 Accuracy results for four medium-imbalanced datasets (Spambase, Ibn Sina, Calif. Housing and Nova). `UNC-CE` outperforms `QBC` on three datasets (((b), (c) and (d)) and loses on Spambase (a). `UNC-IE` loses to `QBC` on all four datasets. F1 results for four relatively skewed datasets (Sick, Zebra, LetterO and Hiva). `UNC-CE` outperforms `QBC` significantly on three datasets ((e), (f) and (h)), and loses on one (g). `UNC-IE` loses to `QBC` on all four datasets.

supported by a small amount of labeled data whereas the lack of evidence is supported by more labeled data. This suggests that the model is more likely to be incorrect in its reasoning that there is a conflict than its reasoning that there is not enough evidence. Further, we showed that the model has higher variance on the `UNC-CE` cases than on the `UNC-IE` cases. Put another way, the model parameters are more "sure" about the uncertainty of the `UNC-IE` cases (lower variance) and therefore the `UNC-IE` cases might indeed continue to be inconclusive even if more labeled data is collected. We compared `UNC-CE` and `UNC-IE` strategies to `QBC` and showed that `UNC-CE` outperforms `QBC` whereas `UNC-IE` loses to `QBC`.

6 Extension to Other Classifiers and Multi-class Classification

In this section, we describe how the evidence-based framework can be extended to other classifiers. We formally define evidence using multinomial naïve Bayes, logistic regression, linear support vector machines, and non-linear support vector machines. Finally, we discuss how it can be generalized to multi-class classification domains.

6.1 Evidence using multinomial naïve Bayes

The probability of a document, $d^{(i)}$, belonging to a class $+1$ is computed using Eq. 16.

$$P(+1|d^{(i)}) = \frac{P(+1) \prod_{1 \leq k^{(i)} \leq n} P(t_k^{(i)} | +1)}{P(d^{(i)})} \qquad (16)$$

where, $t_k^{(i)}$ is the kth term in a document, $d^{(i)}$, $k^{(i)}$ is the number of terms that appear in document, $d^{(i)}$, and n is the dictionary size. A document $d^{(i)}$ can then be classified based on the ratio of $\frac{P(+1|d^{(i)})}{P(-1|d^{(i)})}$:

$$Y = \begin{cases} +1 \ \text{if} \ \left(\frac{P(+1)}{P(-1)} \prod_{1 \leq k^{(i)} \leq n} \frac{P(t_k^{(i)}|+1)}{P(t_k^{(i)}|-1)} \right) > 1 \\ -1 \ \text{otherwise} \end{cases} \qquad (17)$$

From Eq. 17, it follows that the term $t_k^{(i)}$ of document $d^{(i)}$ provides evidence for the positive class if $\frac{P(t_k^{(i)}|+1)}{P(t_k^{(i)}|-1)} > 1$, and it provides evidence for the negative class otherwise. Let $\mathcal{P}_{d^{(i)}}$ and $\mathcal{N}_{d^{(i)}}$ be two sets, such that $\mathcal{P}_{d^{(i)}}$ contains the terms that provide evidence for the positive class and $\mathcal{N}_{d^{(i)}}$ is the set of terms that provide evidence for the negative class:

$$\mathcal{P}_{d^{(i)}} \triangleq \{t_k^{(i)} \mid \frac{P(t_k^{(i)}|+1)}{P(t_k^{(i)}|-1)} > 1\}$$

$$\mathcal{N}_{d^{(i)}} \triangleq \{t_k^{(i)} \mid \frac{P(t_k^{(i)}|-1)}{P(t_k^{(i)}|+1)} > 1\}$$

Then, the total evidence the document, $d^{(i)}$, provides for the positive class is:

$$E_{+1}(d^{(i)}) = \prod_{t_k^{(i)} \in \mathcal{P}_{d^{(i)}}} \frac{P(t_k^{(i)}|+1)}{P(t_k^{(i)}|-1)} \qquad (18)$$

and, the total evidence the document provides for the negative class is:

$$E_{-1}(d^{(i)}) = \prod_{t_k^{(i)} \in \mathcal{N}_{d^{(i)}}} \frac{P(t_k^{(i)}|-1)}{P(t_k^{(i)}|+1)} \qquad (19)$$

6.2 Evidence using Logistic Regression

The parametric model assumed by logistic regression for binary classification is:

$$P(Y = -1|x) = \frac{1}{1 + e^{(w_0 + \sum\limits_{j=1}^{f} w_j x_j^{(i)})}} \tag{20}$$

$$P(Y = +1|x) = \frac{e^{(w_0 + \sum\limits_{j=1}^{f} w_j x_j^{(i)})}}{1 + e^{(w_0 + \sum\limits_{i=1}^{f} w_j x_j^{(i)})}} \tag{21}$$

An instance can then be classified using:

$$Y = \text{sgn}\left(w_0 + \sum_{i=1}^{f} w_j x_j^{(i)}\right) \tag{22}$$

From Eq. 22, it follows that the attribute value $x_j^{(i)}$ of instance $x^{(i)}$ provides evidence for the positive class if $w_j x_j^{(i)} > 0$, and it provides evidence for the negative class otherwise.

Let $\mathcal{P}_{x^{(i)}}$ and $\mathcal{N}_{x^{(i)}}$ be two sets, such that $\mathcal{P}_{x^{(i)}}$ contains the attribute values that provide evidence for the positive class and $\mathcal{N}_{x^{(i)}}$ contains the attribute values that provide evidence for the negative class:

$$\mathcal{P}_{x^{(i)}} \triangleq \{x_j^{(i)} \mid w_j x_j^{(i)} > 0\}$$

$$\mathcal{N}_{x^{(i)}} \triangleq \{x_k^{(i)} \mid w_k x_k^{(i)} < 0\}$$

Then, the total evidence that instance $x^{(i)}$ provides for the positive class is:

$$E_{+1}(x^{(i)}) = \sum_{x_j^{(i)} \in \mathcal{P}_{x^{(i)}}} w_j x_j^{(i)} \tag{23}$$

and, the total evidence that instance $x^{(i)}$ provides for the negative class is:

$$E_{-1}(x^{(i)}) = - \sum_{x_k^{(i)} \in \mathcal{N}_{x^{(i)}}} w_k x_k^{(i)} \tag{24}$$

6.3 Evidence using Linear Support Vector Machines

Support Vector Machines (SVM) maximize the margin of classification:

$$w = \arg\max_{w} \left(y \times (w_0 + \sum_{j=1}^{f} w_j x_j^{(i)})\right) \tag{25}$$

and the classification rule is identical to that of logistic regression (Eqn. 22):

$$Y = \text{sgn}\left(w_0 + \sum_{i=1}^{f} w_j x_j^{(i)}\right) \tag{26}$$

Following the reasoning of evidence using logistic regression, the equations for $E_{+1}(x^{(i)})$ and $E_{-1}(x^{(i)})$ for linear SVM are identical to those for logistic regression.

6.4 Evidence using Non-linear Support Vector Machines

Non-linear SVM maps the data on to a higher dimensional space and uses a linear classifier in a higher dimensional space. For non-linear SVM, the optimization problem is:

$$w = \underset{w}{\arg\min} \lambda \| w \|^2 + \sum_{l=1}^{m} L(w.\phi(x^{(l)}), y^{(l)}) \tag{27}$$

where $w = \sum_{l=1}^{m} \beta_l \phi(x^{(l)})$, $\lambda = \frac{1}{C}$ is the regularization parameter, and $L(y,t) = max(0, 1 - yt)^p$ is a loss function. An instance, $x^{(i)}$, is then classified using:

$$Y = \text{sgn} \sum_{l=1}^{m} \beta_l k(x^{(l)}, x^{(i)}) + b \tag{28}$$

where $k(x^{(l)}, x^{(i)}) = \phi(x^{(l)})^T.\phi(x^{(i)})$ is a kernel function that defines weighted similarity between $x^{(l)}$ and $x^{(i)}$ and β_l is the coefficient which is non-zero for the support vectors and zero for all other instances in the training data.

In case of non-linear SVMs, the evidence that instance $x^{(i)}$ provides for one class or another is it's weighted similarity to the support vectors, $x^{(l)}$, which is defined using a kernel function, $k(x^{(l)}, x^{(i)})$. Let $\mathcal{P}_{x^{(i)}}$ and $\mathcal{N}_{x^{(i)}}$ be two sets for instance $x^{(i)}$, such that $\mathcal{P}_{x^{(i)}}$ contains the support vectors that provide evidence for the positive class for $x^{(i)}$ and $\mathcal{N}_{x^{(i)}}$ contains the support vectors that provide evidence for the negative class for $x^{(i)}$:

$$\mathcal{P}_{x^{(i)}} \triangleq \{x^{(j)} \mid \beta_j k(x^{(j)}, x^{(i)}) > 0\}$$

$$\mathcal{N}_{x^{(i)}} \triangleq \{x^{(k)} \mid \beta_k k(x^{(k)}, x^{(i)}) < 0\}$$

Then, the total evidence that instance $x^{(i)}$ contains for the positive class is:

$$E_{+1}(x^{(i)}) = \sum_{x^{(j)} \in \mathcal{P}_{x^{(i)}}} \beta_j k(x^{(j)}, x^{(i)}) \tag{29}$$

and, the total evidence that instance $x^{(i)}$ contains for the negative class is:

$$E_{-1}(x^{(i)}) = \sum_{x^{(k)} \in \mathcal{N}_{x^{(i)}}} \beta_k k(x^{(k)}, x^{(i)}) \tag{30}$$

6.5 Evidence for Multi-class Classification

For binary classification, all three types of uncertainties (Equations 1, 2, and 3) prefer instances closest to the decision boundary as specified by Equations 5, 22, and 26. However, their preferences differ in multi-class classification. The entropy approach (Eq. 1), for example, considers overall uncertainty and takes into account all classes, whereas the maximum conditional approach (Eq. 2) considers how confident the model is about the most likely class. To keep the discussion simple and brief, and as a proof-of-concept, we show how the evidence for multi-class can be extended for naive Bayes (Eq. 4) when used with the margin uncertainty approach (Eq. 3).

The margin uncertainty prefers instances for which the difference between the probabilities of most-likely class y_m and next-likely class y_n is minimum. Let $\mathcal{M}_{x^{(i)}}$ and $\mathcal{N}_{x^{(i)}}$ be two sets, such that $\mathcal{M}_{x^{(i)}}$

contains the attribute values that provide evidence for the most-likely class and $\mathcal{N}_{x^{(i)}}$ contains the attribute values that provide evidence for the next likely class:

$$\mathcal{M}_{x^{(i)}} \triangleq \{x_j^{(i)} \mid \frac{P(x_j^{(i)}|y_m)}{P(x_j^{(i)}|y_n)} > 1\}$$

$$\mathcal{N}_{x^{(i)}} \triangleq \{x_k^{(i)} \mid \frac{P(x_k^{(i)}|y_n)}{P(x_k^{(i)}|y_m)} > 1\}$$

Then, the total evidence that instance $x^{(i)}$ provides for the most-likely class (in comparison to the next-likely class) is:

$$E_m(x^{(i)}) = \prod_{x_j^{(i)} \in \mathcal{M}_x^{(i)}} \frac{P(x_j^{(i)}|y_m)}{P(x_j^{(i)}|y_n)} \tag{31}$$

and, the total evidence that instance $x^{(i)}$ provides for the next-likely class (in comparison to the most-likely class) is:

$$E_n(x^{(i)}) = \prod_{x_k^{(i)} \in \mathcal{N}_x^{(i)}} \frac{P(x_k^{(i)}|y_n)}{P(x_k^{(i)}|y_m)} \tag{32}$$

7 Conclusion and Future Work

We introduced an evidence-based framework that can uncover the reasons for a model's uncertainty. We used this framework to distinguish between two types of uncertainties: a model is uncertain about an instance due to strong and conflicting evidence for both classes (conflicting-evidence uncertainty) versus a model is uncertain because it does not have sufficient evidence for either class (insufficient-evidence uncertainty). Traditional uncertainty sampling does not distinguish between these types of uncertainties, but our empirical evaluations showed that making this distinction has a big impact on the performance of learner: while insufficient-evidence uncertain instances provided the least value to an active learner, actively labeling conflicting-evidence uncertain instances significantly improved the learning efficiency.

We provided analytical and empirical results showing that the conflicting-evidence instances are underrepresented in the labeled data compared to the insufficient-evidence instances. We further provided empirical results showing that the model has higher variance on the conflicting-evidence instances than on the insufficient-evidence instances. These two results suggest that the model is more likely to be incorrect in its decision that there is a conflict than its decision that the case is inconclusive.

We primarily focused on evaluating the performance of evidence-based framework for naïve Bayes classifier. We provided empirical evaluations on several real-world datasets and provided the analysis of the evidence-based framework using naïve Bayes on binary classification tasks. Evaluating our methods using other classifiers such as logistic regression and support vector machines, and on multiclass classification tasks is left as future work. Even though we observed on most datasets that UNC-CE outperforms UNC-1 and UNC-IE was even worse than UNC-t, there was one case (Hiva-AUC) for which UNC-IE outperforms UNC-1. The current framework cannot predict which of the two uncertainties would help the learning the most. The evidence-based framework is not specific to improving any particular performance measure and is still a decision boundary approach.

We combined the evidence-based framework and uncertainty sampling using a two-step approach in Sect. 2.3.2: we first ranked the instances according to uncertainty and then applied the evidence-based

framework on the top k instances. We observed that this simple approach worked well, but it requires us to define what is uncertain, e.g. defining a threshold for number of top uncertain instances. In future, we would like to investigate multi-criteria optimization approaches (Steuer, 1989) for combining uncertainty sampling and the evidence-based framework.

It is well-known in the active learning community that uncertainty sampling is susceptible to noise and outliers (Settles and Craven, 2008). `UNC-CE` prefers instances that have low density with respect to training data. We showed that `UNC-CE` performs quite well on many datasets. However, we do not know whether preferring instances selected by `UNC-CE` over instances selected by `UNC-1` and `UNC-IE` makes it more susceptible to noise and outliers. This needs to be verified using controlled experiments with synthetic datasets.

Another interesting future work is to utilize the formalism of (Senge et al, 2014) to investigate whether any parallels and similarities can be drawn between conflicting versus insufficient-evidence and aleatoric versus epistemic uncertainty cases. In this article we have distinguished between the two types of uncertainties and have evaluated its benefit in selecting informative instances for active learning for classification. The proposed framework may be applied to regression problems as well, but its discussion is beyond the scope of this article.

Acknowledgments

References

Abe N, Mamitsuka H (1998) Query learning strategies using boosting and bagging. In: Proceedings of the Fifteenth International Conference on Machine Learning, pp 1–9

Bilgic M, Mihalkova L, Getoor L (2010) Active learning for networked data. In: Proceedings of the 27th International Conference on Machine Learning, pp 79–86

Chao C, Cakmak M, Thomaz AL (2010) Transparent active learning for robots. In: 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, pp 317–324

Cohn DA (1997) Minimizing statistical bias with queries. In: Advances in Neural Information Processing Systems, pp 417–423

Cohn DA, Ghahramani Z, Jordan MI (1996) Active learning with statistical models. Journal of Artificial Intelligence Research 4:129–145

Dagan I, Engelson SP (1995) Committee-based sampling for training probabilistic classifiers. In: Proceedings of the Twelfth International Conference on Machine Learning, pp 150–157

Donmez P, Carbonell JG, Bennett PN (2007) Dual strategy active learning. In: Machine Learning: ECML 2007, Springer, pp 116–127

Frank A, Asuncion A (2010) UCI machine learning repository. URL http://archive.ics.uci.edu/ml

Frey PW, Slate DJ (1991) Letter recognition using holland-style adaptive classifiers. Machine Learning 6(2):161–182

Gu Q, Zhang T, Han J, Ding CH (2012) Selective labeling via error bound minimization. In: Advances in Neural Information Processing Systems, pp 323–331

Gu Q, Zhang T, Han J (2014) Batch-mode active learning via error bound minimization. In: Proceedings of the Thirtieth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-14), AUAI Press, Corvallis, Oregon, pp 300–309

Guyon I, et al (2011) Datasets of the active learning challenge. Journal of Machine Learning Research

Hoi SC, Jin R, Lyu MR (2006a) Large-scale text categorization by batch mode active learning. In: Proceedings of the 15th International Conference on World Wide Web, ACM, pp 633–642

Hoi SC, Jin R, Zhu J, Lyu MR (2006b) Batch mode active learning and its application to medical image classification. In: Proceedings of the 23rd International Conference on Machine learning, ACM, pp 417–424

Lewis DD, Gale WA (1994) A sequential algorithm for training text classifiers. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Springer-Verlag New York, Inc., pp 3–12

Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics, pp 142–150

MacKay DJ (1992) Information-based objective functions for active data selection. Neural Computation 4(4):590–604

McCallum A, Nigam K, et al (1998) A comparison of event models for naive bayes text classification. In: AAAI-98 Workshop on Learning for Text Categorization, Citeseer, vol 752, pp 41–48

Melville P, Mooney RJ (2004) Diverse ensembles for active learning. In: Proceedings of the Twenty-first International Conference on Machine Learning, pp 74–

Mitchell TM (1982) Generalization as search. Artificial Intelligence 18(2):203–226

Nguyen HT, Smeulders A (2004) Active learning using pre-clustering. In: Proceedings of the Twenty-first International Conference on Machine learning, ACM, p 79

Pace RK, Barry R (1997) Sparse spatial autoregressions. Statistics & Probability Letters 33(3):291–297

Roy N, McCallum A (2001) Toward optimal active learning through sampling estimation of error reduction. In: Proceedings of the Eighteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., ICML '01, pp 441–448

Sculley D (2007) Online active learning methods for fast label-efficient spam filtering. In: Fourth Conference on Email and Anti-Spam (CEAS)

Segal R, Markowitz T, Arnold W (2006) Fast uncertainty sampling for labeling large e-mail corpora. In: Third Conference on Email and Anti-Spam (CEAS)

Senge R, Bösner S, Dembczyński K, Haasenritter J, Hirsch O, Donner-Banzhoff N, Hüllermeier E (2014) Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. Information Sciences 255:16–29

Settles B (2012) Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 6(1):1–114

Settles B, Craven M (2008) An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp 1070–1079

Seung HS, Opper M, Sompolinsky H (1992) Query by committee. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, ACM, pp 287–294

Sharma M, Bilgic M (2013) Most-surely vs. least-surely uncertain. In: IEEE 13th International Conference on Data Mining (ICDM), pp 667–676

Sindhwani V, Melville P, Lawrence RD (2009) Uncertainty sampling and transductive experimental design for active dual supervision. In: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, pp 953–960

Steuer RE (1989) Multiple Criteria Optimization: Theory, Computations, and Application. Krieger Pub Co

Thompson CA, Califf ME, Mooney RJ (1999) Active learning for natural language parsing and information extraction. In: Proceedings of the Sixteenth International Conference on Machine Learning, pp 406–414

Tong S, Chang E (2001) Support vector machine active learning for image retrieval. In: Proceedings of the Ninth ACM International Conference on Multimedia, ACM, pp 107–118

Xu Z, Yu K, Tresp V, Xu X, Wang J (2003) Representative sampling for text classification using support vector machines. In: Advances in Information Retrieval, Lecture Notes in Computer Science - Volume 2633, 2003, pp 393–407

Yu K, Bi J, Tresp V (2006) Active learning via transductive experimental design. In: Proceedings of the 23rd International Conference on Machine Learning, ACM, pp 1081–1088

Zhang C, Chen T (2002) An active learning framework for content-based information retrieval. IEEE Transactions on Multimedia 4(2):260–268

Zhu J, Wang H, Yao T, Tsou BK (2008) Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In: Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1, pp 1137–1144