Evidential Mixture Machines: Deciphering Multi-Label Correlations for Active Learning Sensitivity

Anonymous Author(s)

ABSTRACT

Multi-label active learning is an essential and challenging aspect of contemporary machine learning, often hindered by the complexities of managing expansive and sparse label spaces. This challenge is intensified in active learning scenarios where labeling resources are limited. Drawing inspiration from existing mixture of Bernoulli models, which effectively compress the label space into a more manageable weight coefficient space through the learning of correlated Bernoulli components, we introduce a novel evidential mixture machines (EMM) model. This model advances uncertainty-aware learning from input features to the predicted coefficients and components. It leverages mixture components obtained through unsupervised learning, enhancing prediction accuracy by learning to predict coefficients evidentially and aggregating component offset predictions as proxy pseudo counts. Furthermore, our approach employs evidential uncertainty combined with predicted label embedding covariances for active sample selection, leading to a multi-source uncertainty metric that is richer than simple uncertainty scores. Experiments on synthetic data demonstrate the effectiveness of evidential uncertainty prediction and capturing the label correlation using predicted components and experiments on real-world datasets show improved performance over existing multi-label active learning methods.

ACM Reference Format:

Anonymous Author(s). 2024. Evidential Mixture Machines: Deciphering Multi-Label Correlations for Active Learning Sensitivity. In . ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In recent years, machine learning methods, especially deep learning models, have seen great success in various tasks, prominently classification tasks. Nevertheless, such success is often coupled with the demand for a large amount of labeled data samples for model training, which might not be readily available in knowledge-rich domains. Active learning (AL) is a paradigm where we have access to abundant unlabeled data instances and a limited labeling budget [31] [14]. Most AL methods focus on the selection strategy that helps the machine learner achieve better performances with an informed selection of labeled data instances. However, while AL for standard classification tasks has been studied extensively, an important task that is multi-label classification has been overlooked by many.

In real-world problems, each data instance can be easily associated with more than one label [10, 23, 34]. For machine learning models, the difference between the multi-class problem where only one ground truth label is associated with a data instance and the multi-label classfication (MLC) problem is fundamental. On one hand, although we can transform the MLC problem into multiple binary problems, we lose the common underlying correlations between input features and labels by doing so. There are also other challenges that come with the transformed binary problems: the number of classifiers may be large due to the large label space, thus the separated training process can be costly. Also, many labels are relatively rare and may depend on other labels, and we can not learn such dependencies when we isolate these labels. On the other hand, when we build a joint learning problem for all labels, the biggest challenge is to combine common labels with rare labels in a balanced way. For the rare labels, there might be few positive data instances. A typical model is likely to optimize the feature embedding according to popular labels or adopt the so-called "shortcut" learning, making a direct connection from the input features to the rare labels more difficult. Although we now have access to common and rare labels at the same time, it might not suffice to promote the learning of their correlations if we directly model the labels from the input features.

These unique challenges of the MLC problem become more pronounced in AL settings, where the rare labels become even more scarce. They also increase the difficulty of obtaining high-quality uncertainty estimation, which is often crucial in many AL selection strategies as it allows us to know when the model "does not know" in order to make an educated decision on which data instances to label. To address these challenges, we turn to the mixture of Bernoulli model, which can model a large label space with a small number of mixture components. There are existing methods that try to capture the label correlation using such a model [15, 26]. However, to connect with the input features, they either resort to a purely conditional case, namely conditional Bernoulli mixtures (CBM) [15] or use a conjugate classification head, which is a Gaussian Process model (GP-B^{2}M) [26]. For the former, a distinct set of label clusters is predicted for each data instance, meaning that the label correlations are completely separated from the learning tasks. This shortcome makes the CBM model unsuitable for AL. The latter relies on the Gaussian Process (GP) which outputs the weights of the mixture components. For the intended task, a complete GP is too expensive, while a sparse GP has limited predictive ability. Unlike the CBM model, GP-B^{2}M creates a set of global label clusters. However, the label prediction for rare labels remains challenging because they can only ever be as good as the best label cluster available. Furthermore, the uncertainty quantification is superficial because it only captures the approximate covariance of the label

Conference'24, August 2024, Barcelona, Spain

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

prediction and the variance of the GP predictions. This point estimate does not fully capture the unknown and is not sufficiently effective in AL.

In this work, we propose to combine the mixture of Bernoulli with a deep evidential model, both incorporating deep learning and enabling more fine-grained uncertainty analysis. The deep learning model needs only a forward pass during prediction time, which is much more efficient than estimating the predictive distribution in a random processes model. The uncertainty analysis can now include each evidential prediction. This flexibility is crucial in multi-label AL as the informativeness of data samples can be more complicated than samples with one-hot labels. The proposed model is composed of a shared encoder that provides embeddings and two decoders that predict the weight coefficients and the proxy pseudo counts correction for the Posterior Beta which models the global label cluster, respectively. The weight coefficient predictor is trained as a deep evidential regression model, which makes uncertainty-aware predictions of optimal assignments to the mixture of label clusters for each data instance. By using a deep evidential model as the classification head, we largely improve the efficiency during inference time as we do not need to approximate the predictive distribution of an entire statistical model. By sharing the encoder, we maintain a close connection beginning from the input features, through the weight coefficients, and to the label clusters. Compared to the CBM model, the proposed Evidential Mixture Machines (EMM) maintains a global label cluster. Compared to the GP-B^{2}M model, the EMM model allows adjustment to the label clusters based on each data instance. Additionally, the uncertainty information can be captured by both the fine-grained uncertainty decomposition from the evidential posterior parameters and the predicted discrepancy between the global label cluster and the proxy pseudo counts. In the evidential formulation, we provide a conjugate interpretation of the Normal Inverse Gamma parameters, giving these parameters a more impactful pseudo count meaning and obtaining a novel overall evidence quantification. In the training process, we propose to follow the evidential training step with a joint training step that empowers the model to correct the global label clusters for the sake of predicting labels for each individual data sample. The joint training step effectively improves the ability to make positive predictions for rare labels and consequently improves the label-perspective uncertainty quantification that benefits active sample selection.

With both the demonstration on synthetic datasets and the AL evaluation on real-world datasets, we show the effectiveness of the EMM model. Our main contribution can be summarized as:

- we propose a novel integration of deep evidential learning with multi-label classification.
- We propose advanced uncertainty quantification in the integrated model for active learning.
- We empirically validated that the proposed model improves the performance for large label spaces, especially on sparse and rare labels.

2 RELATED WORKS

In the realm of multi-label classification (MLC) [36] [23], Binary Relevance Models (BRMs) have gained widespread use, leading to the development of various active learning (AL) models based on BRMs. Notable examples include employing the estimated reduction of a BRM loss function as an uncertainty criterion for data sampling, as demonstrated by [33].

A large portion of the multi-label active learning work do not require annotators to label all possible labels for a given data instance [6, 21, 32]. The main consideration behind these approaches is the significant reduction in annotator labeling costs. However, this strategy inevitably breaks the inherent connections between labels, making it impossible to comprehensively measure the informativeness of a data instance using label correlations. *[dayou: justify]*Additionally, models designed to handle partially labeled data are required, limiting the applicability of such methods. Therefore, in this paper, we will not compare our approach to these methods.

Other approaches integrate the properties of the support vectors of individual support vector machines (SVM) within BRMs, using label correlation more as a means to simplify the querying process than to enhance active sampling [7–9, 25, 28]. While these methods incorporate label correlation to some extent, such as through label inconsistency [16], label ranking [22], or learning reularization[35], they do not systematically capture label correlations, potentially leading to imprecise uncertainty measures in ML-AL contexts.

Some existing models attempt to explicitly capture label correlations or use latent embeddings to facilitate active multi-label sampling. For instance, the CBM model uses the approximate entropy of predicted labels for data sampling [5], but its dependency on an external multi-class classifier for predicting component coefficients complicates AL due to the challenges in model selection and parameter tuning. Furthermore, CBM, designed primarily for MLC rather than AL, predicts distinct label clusters for each data sample without discovering global label clusters, thus limiting its effectiveness in multi-label AL [15]. Other approaches like correlation-aware method for transfer learning [6] struggle with large and sparse label spaces due to their reliance on kernel functions for measuring label similarity. Compressed sensing (CS) techniques [27, 30] innovative in learning latent embeddings to capture label correlations but assume labels are drawn from a Gaussian distribution and are not efficient in AL, especially in early stages with limited training data. In [26], a Bayesian mixture of Bernoulli model is proposed. A set of global label clusters are captured in a Bayesian manner. However, the inference process of the model is complicated and the fixed label clusters limit the predictive ability in the final label space.

Evidential models have been developed to enable fine-grained uncertainty quantification in deep learning (DL) models [24] [1]. These models introduce a higher order conjugate prior distribution over the likelihood distribution, and train the DL model to output the parameters of the higher order distribution [4] [3]. The higher order distribution enables the model to express the fine-grained uncertainty information. Evidential models have been successfully extended to classification [11, 17], regression [1, 18], action recognition [2], OOD detection [12], and meta-learning problems [19]. We extend the evidential deep learning framework to our setting that leads to novel fine-grained uncertainty guided active-learning for multi-label classification.

3 METHODOLOGY

3.1 Problem Setting

In our multi-label active learning problem, the essential task is to predict a multi-variate 0-1 label vector $\mathbf{y} = (y_1, ..., y_L)^\top \in \{0,1\}^L$ from the input features $\mathbf{x} \in \mathbb{R}^M$. For AL, we start with a small initial labeled set $\mathcal{S}_L$, and a large unlabeled pool $\mathcal{S}_U$ ($N_L = |\mathcal{S}_L| << N_U = |\mathcal{S}_U|$). The AL strategy $\mathcal{A}(\mathbf{x})$ selects the instance $\mathbf{x}^*$ from $\mathcal{S}_U$ to be labeled as a batch b_t and added to $\mathcal{S}_L$.

3.2 Evidential Mixture Machines

Compared to existing methods, our novel contribution to model learning lies in how we connect label clusters to input features, enable evidential uncertainty analysis for weight coefficient predictions, and learn the final labels in a joint manner. Then, we will utilize these model properties for AL in the next subsection.

3.2.1 Mixture of Bernoulli. The set of label clusters contains K mixture components. Each mixture component is a L-variate Bernoulli distribution $\mathbf{z}_k = \prod_{l=1}^{L} \text{Bernoulli}(y_l, \mu_{kl})$ that captures a type of label distribution, where $\text{Bernoulli}(\mu_{kl}) = \mu_{kl}^{y_l}(1-\mu_{kl})^{(1-y_l)}$ and L is the total number of labels. The Bernoulli parameter μ_{kl} has the conjugate prior $\text{Beta}(a_{kl}, b_{kl})$. Initially, the mixture of Bernoulli can be found through label-only learning. Using an EM algorithm, we can learn the initial components $\boldsymbol{\mu}^{(0)}_{K\times L}$ from the labeled samples $\mathcal{S}_L$. These components can model the set of labels $p(\mathbf{y}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \pi_k \prod_{l=1}^{L} \text{Bernoulli}(y_l, \mu_{kl})$, where π_k is the weight coefficient for component k. However, since this is a label-only process, we are missing the connection to the input features $\mathbf{x}$.

3.2.2 Connecting to Input Features. For a conditional model such as CBM, the connection is through $\pi_k = p(\mathbf{z}_k|\mathbf{x})$ and $\mu = \mu(\mathbf{x})$. The model is still trained in an EM manner so that the predictions $\hat{\mathbf{y}}$ can be made for each $\mathbf{x}$. For a conjugate model such as GP-B^{2}M, the connection is through $\pi_k = p(\mathbf{z}_k|\text{GP}_k(\mathbf{x}))$, and the variational training process of the conjugate model impacting the posterior $\boldsymbol{\mu}$. As mentioned before, one issue with CBM is the disconnected prediction models of $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$, while the inference process of GP-B^{2}M is too expensive.

In our proposed model, we also predict both $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\mu}}$ from $\mathbf{x}$ using two decoder networks $g_{\boldsymbol{\pi}}(\cdot)$ and $g_{\boldsymbol{\mu}}(\cdot)$. However, we maintain the connection by using a shared encoder network $e(\mathbf{x})$. The structure is shown in Figure 1. Specifically, $g_{\boldsymbol{\pi}}(e(\mathbf{x}))$ is an evidential model that predicts $\boldsymbol{\pi}$ along with its evidence parameters. Furthermore, instead of directly letting the network predict $\boldsymbol{\mu}$ from $g_{\boldsymbol{\mu}}(\cdot)$, we make $g_{\boldsymbol{\mu}}(\cdot)$ output proxy pseudo counts $\hat{a}_{kl}$ and $\hat{b}_{kl}$ to be combined with the initial $(\boldsymbol{\mu}^{(0)}_{K\times L}|a^{(0)}_{kl}, b^{(0)}_{kl})$. This allows us to have a global set of label clusters as in GP-B^{2}M, without actually implementing the entire conjugate Bayesian model.

3.2.3 Evidential Weight Coefficient Predictor. We want the coefficient predictor to output not only the predicted value of the coefficients but also the confidence in the prediction. Counterintuitively, even though $\sum_{k=1}^{K} \pi_k = 1$, this is a regression problem instead of a classification problem. This is because we do not want the model to output a single highly confident prediction entry π_k for each instance, but a combination of k values that best reconstruct the final label predictions. Thus, this branch needs to be an evidential regression model.

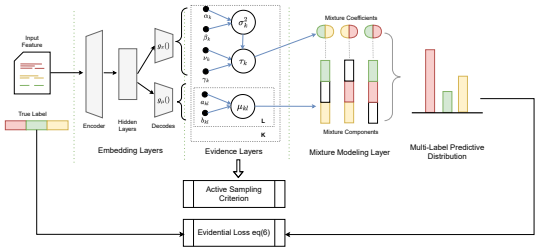


Figure 1: Overall framework of EMM.

To this end, we place a higher-order Normal Inverse Gamma (NIG) prior ($\text{NIG}(\tau, \sigma^2|\mathbf{p}) = \mathcal{N}(\tau|\gamma, \frac{\sigma^2}{v})\Gamma^{-1}(\sigma^2|\alpha, \beta)$) over the regression model's gaussian output $\mathcal{N}(\pi|\tau, \sigma)$. The evidential model is trained to output the NIG parameters $\mathbf{p} = (\gamma, v, \alpha, \beta)$ similar to [1]. In this evidential model, the gaussian likelihood interacts with the NIG prior leading to a a Student-t predictive distribution:

$$
\begin{aligned}
p(\pi|x,\mathbf{p}) &= \int_{\tau}\int_{\sigma^2} p(\pi|x,\tau,\sigma^2)\text{NIG}(\tau,\sigma^2|\mathbf{p})\mathrm{d}\tau\mathrm{d}\sigma^2 \\
&= \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)}\sqrt{\frac{v}{2\pi\beta(1+v)}}\Big(1+\frac{v(\pi-\gamma)^2}{2\beta(1+v)}\Big)^{-(\alpha+\frac{1}{2})} \\
&= \text{St}\Big(\pi;\gamma,\frac{\beta(1+v)}{v\alpha},2\alpha\Big) \qquad (1)
\end{aligned}
$$

Here, the evidential model predicts coefficients for input x as:

$$
\hat{\pi} = \mathbb{E}_{p(\pi|x,\mathbf{p})}[\pi] = \int \pi p(\pi|x,\mathbf{p})\mathrm{d}\pi = \gamma \qquad (2)
$$

Where the predictor branch $g_{\boldsymbol{\pi}}(e(\mathbf{x}))$ outputs the NIG parameters γ, v, α, β. The evidential model, through its higher order NIG prior, can quantify the aleatoric (AL) and epistemic (EP) uncertainty [1] as $\text{AL} = \mathbb{E}[\sigma^2] = \frac{\beta}{\alpha-1}$,$\text{EP} = \text{Var}[\mu] = \frac{\beta}{v(\alpha-1)}$. In this evidential framework, due to the conjugacy of the NIG prior with the Gaussian likelihood, the posterior is also the NIG distribution. Moreover, in this model, after interacting with N i.i.d. data points $(\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_N)$, the posterior NIG parameters update [20] as

$$
v_N = v + N, \quad \gamma_N = \frac{v}{v_N}\gamma + \frac{1}{v_N}\sum_{n=1}^{N}\pi_n, \quad \alpha_N = \alpha + \frac{N}{2} \qquad (3)
$$

$$
\beta_N = \beta + \frac{1}{2}\sum_{k=1}^{N}(\pi_n - \sum_{n=1}^{N}\frac{\pi_n}{N})^2 + \frac{Nv}{2(v+N)}(\sum_{n=1}^{N}\frac{\pi_n}{N} - \gamma)^2 \qquad (4)
$$

As can be seen, each observation impacts the confidence of the model through v, α, and β. Thus, the three evidential parameters v, α, and β contribute to model's evidence/confidence. v and α can be seen as the pseudo-counts and directly impact the model evidence i.e., quantify the confidence on the prior mean and the prediction of a target data sample, respectively. Furthermore, a large β leads to a low confidence in the model's prediction, which implies a lack of evidence. Additionally, each observation increases the pseudo-count of v by 1, and α by $\frac{1}{2}$. We integrate all these insights to

quantify the overall evidence $\mathcal{E}$ of the model as

$$\mathcal{E} = v + \frac{1}{2}\alpha + \frac{1}{\beta} \tag{5}$$

We train the model to maximize The likelihood under the predictive Student-t distribution. With this, the loss is given as:

$$\mathcal{L}_{NLL} = -\log(p(\pi_k|x,\mathbf{p}) = \frac{1}{2}\log(\frac{\pi}{v}) - \alpha\log(\Omega) + (\alpha+\frac{1}{2})\log((\pi_k-\gamma)^2 v + \Omega) + \log(\frac{\Gamma(\alpha)}{\Gamma(\alpha+\frac{1}{2})})$$

where $\Omega = 2\beta(1+v)$. Additionally, we want the model's confidence/evidence for the prediction to be low when the prediction is incorrect. To this end, we introduce a evidence-based regularization

$$\mathcal{L}_{REG} = (\pi_k - \gamma)^2 \cdot \mathcal{E}$$

The regularization penalizes the highly confident wrong predictions, and ensures model's confidence is rightly placed. The overall loss is given by

$$\mathcal{L}_{EVID} = \mathcal{L}_{NLL} + \lambda_{reg}\mathcal{L}_{REG} \tag{6}$$

Where λ controls the effect of the regularization to the model training.

The integration of an evidential model in our multi-label classification approach presents several distinct advantages, particularly in addressing the inherent complexities of active learning (AL) environments. Firstly, evidential models provide a more nuanced and sophisticated mechanism for uncertainty quantification. This is crucial in AL settings, especially when dealing with sparse and rare labels, where traditional models often struggle. By effectively capturing and quantifying uncertainty, our approach allows for more informed and strategic decisions in the selection of data instances for labeling, optimizing the use of limited labeling resources. Furthermore, the evidential model facilitates a deeper understanding of the underlying label correlations, enabling the model to make more accurate predictions across a broad spectrum of labels, including those that are less frequent. This leads to a significant improvement in the overall performance of the classifier, especially in scenarios where conventional methods might overlook subtle but crucial label dependencies. Additionally, the evidential approach inherently enhances the interpretability of the model's predictions, offering insights into the confidence and reliability of these predictions. This aspect is particularly valuable in knowledge-rich domains where understanding the model's decision-making process is as important as the accuracy of the predictions themselves. Thus, the incorporation of an evidential model into our multi-label classification framework marks a substantial advancement, offering a robust, efficient, and insightful solution to the challenges posed by large and complex label spaces in active learning scenarios.

In one learning round, we start with training this branch to fit the set of $\Pi^{(0)}$ optimized for the initial $\boldsymbol{\mu}^{(0)}_{K\times L}$. Then, we move on to the joint training stage where we alternate between training the coefficient predictor and the full model to fit the labels $\mathbf{y}$.

3.2.4 Joint Multi-label Training with Label Clusters. Once we have the coefficient predictor branch, we move on to training the full model in order to make the final label predictions. We first freeze $e(\cdot)$ and $g_{\boldsymbol{\pi}_k}(\cdot)$ to train $g_{\boldsymbol{\mu}}(\cdot)$. The network outputs of dimension $2\cdot N\cdot K\cdot L$ are split into $\hat{a}$ and $\hat{b}$ and added to $a^{(0)}$ and $b^{(0)}$ in a weighted fashion: $a_{kl}(\mathbf{x}) = a^{(0)}_{kl} + w_\mu \hat{a}_{kl}(\mathbf{x}), b_{kl}(\mathbf{x}) = b^{(0)}_{kl} + w_\mu \hat{b}_{kl}(\mathbf{x})$. The new Bernoulli parameter for each instance is then computed by $\mu_{kl}(\mathbf{x}) = a_{kl}(\mathbf{x})/(a_{kl}(\mathbf{x}) + b_{kl}(\mathbf{x}))$. The label prediction is $\hat{y}_l(\mathbf{x}) = \sum_k \pi(\mathbf{x})\mu_{kl}(\mathbf{x})$. The model is trained using a soft margin multi-label loss:

$$\mathcal{L}_{SoftMargin} = -\frac{1}{L}\sum_{l=1}^{L} y_l \log(\frac{1}{1+\exp(-\hat{y}_l)}) + (1-y_l)\log(\frac{\exp(-\hat{y}_l)}{1+\exp(-\hat{y}_l)}) \tag{7}$$

We then adopt the evidential pseudo-count style update of the label clusters, we would have $a^{(1)}_{kl} = a^{(0)}_{kl} + \sum_{k=1}^{K} \hat{\pi_k}(\mathbf{x}_n) y_{nl}$ and $b^{(1)}_{kl} = b^{(0)}_{kl} + \sum_{k=1}^{K} \hat{\pi_k}(\mathbf{x}_n)(1-y_{nl})$. However, this update only makes the correction based on the predicted weights$\hat{\mu_k}$. If we keep updating in this way, the biases build up and the popular labels will be dominant in future components. Thus, we also include a weighted update based on the predicted proxy pseudo-counts, similar to when we make label predictions: $a_{kl}(\mathbf{x}) = a^{(0)}_{kl} + w_{up}\hat{a}_{kl}(\mathbf{x}), b_{kl}(\mathbf{x}) = b^{(0)}_{kl} + w_{up}\hat{b}_{kl}(\mathbf{x})$. This step ensures that our model mutually benefits from the coefficient predictor and the proxy pseudo-count predictor. The initial components are learned unsupervised and do not make up for the training of the predictors. By introducing the joint update, we connect the two predictors more closely.

The joint training step of EMM addresses an important problem with the mixture model formulation, where the model prediction is restricted by the mixture components $\boldsymbol{\mu}$. Because the weight coefficients $0 \le \pi_k \le 1$, the label prediction for y_l can only be as great as $\max_k \mu_{kl}$. If we only make updates to the mixture components using $a^{(1)}_{kl} = a^{(0)}_{kl} + \sum_{k=1}^{K} \hat{\pi_k}(\mathbf{x}_n) y_{nl}$ and $b^{(1)}_{kl} = b^{(0)}_{kl} + \sum_{k=1}^{K} \hat{\pi_k}(\mathbf{x}_n)(1-y_{nl})$, the rare labels will still suffer from the label imbalance which will be reflected in $\max_k \mu_{kl} = \max_k \frac{a_{kl}}{a_{kl}+b_{kl}}$. By incorporating the joint training, we make the prediction based on $a_{kl}(\mathbf{x}) = a^{(0)}_{kl} + w_\mu \hat{a}_{kl}(\mathbf{x}), b_{kl}(\mathbf{x}) = b^{(0)}_{kl} + w_\mu \hat{b}_{kl}(\mathbf{x})$, allowing the model to better fit the labels using instance-wise predictions $\hat{a}_{kl}(\mathbf{x}), \hat{b}_{kl}(\mathbf{x})$. The soft margin multi-label loss effectively brings the benefits of binary relevance machines into model training because it promotes positive predictions through the multi-label one-versus-rest formulation.

At this stage, we have combined the advantages of Bayesian mixture models, deep evidential models, and a bi-level multi-label problem formulation to obtain a powerful multi-label classification model. Next, we introduce how the evidential flavor can provide fine-grained uncertainty information and benefit active learning.

3.3 Active Learning Strategy

In order to select the most informative samples given the particularly small initial budget, we adopt an uncertainty-oriented selection strategy. From the proposed EMM model, we can obtain uncertainty information from three sources: weight coefficient branch, proxy pseudo count predictor, and the final label predictions.

The first part of the uncertainty information is from the evidential model that predicts the weight coefficients. The evidential model naturally decomposites the aleatoric uncertainty $\mathbb{E}[\sigma^2_{\pi_k}]$ and

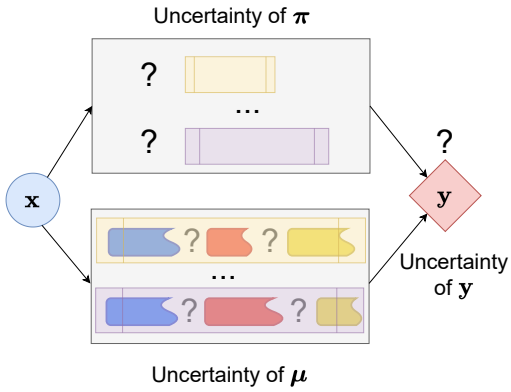


Figure 2: A visualization of the multi-source uncertainty composition

the epistemic uncertainty $\text{Var}[\pi_k] = \mathbb{E}[\sigma^2_{\pi_k}]/\nu$. For AL purposes, we should target the samples that give us the most epistemic uncertainty, which can potentially improve the model's knowledge of the unknown. Thus, the first part of the selection function is $\mathcal{A}_{\pi_k}(\mathbf{x}) = \frac{\beta_k(\mathbf{x})}{\nu_k(\mathbf{x})(\alpha_k(\mathbf{x})-1)}$. From the weight coefficient perspective, this criterion searches for the least confident samples based on our current model. Selecting these samples will help us quickly gain knowledge of the connection between input features and the weight coefficients, which is the determining factor for predictive performance.

The second part of the uncertainty information comes from the proxy pseudo counts. We can compare the current components and the updated components when the proxy pseudo counts for $\mathbf{x}$ are added, and select the samples that would introduce more difference to the current model: $\mathcal{A}_{\mu}(\mathbf{x}) = -\text{CosineSimilarity}(\boldsymbol{\mu}, \boldsymbol{\mu}'(\mathbf{x})) = \frac{\boldsymbol{\mu}\cdot\boldsymbol{\mu}'(\mathbf{x})}{\|\boldsymbol{\mu}\|\cdot\|\boldsymbol{\mu}'(\mathbf{x})\|}$. Because the label clusters play the most important role in recreating the label space, we should try to capture as much latent label correlation as possible. This requires sufficient exploration in the label space. Selection criterion $\mathcal{A}_{\mu}(\mathbf{x})$ does exactly this by searching for data samples that are potentially the most different from the currently captured label space.

The last part of the uncertainty information is computed over the final label prediction. Since the full model is a mixture of Bernoulli, we can compute the expected covariance of the predicted label distribution. Here, we adopt the point estimate in the same way as existing methods by:

$$\text{cov}[\hat{\mathbf{y}}|\mathbf{x}] = \sum_k \pi_k(\text{diag}(\boldsymbol{\mu}(1-\boldsymbol{\mu})) + \boldsymbol{\mu}_k\boldsymbol{\mu}_k^\top) - \mathbf{p}(\hat{\mathbf{y}}|\mathbf{x})\mathbf{p}(\hat{\mathbf{y}}|\mathbf{x})^\top \quad (8)$$

Note that ideally we would like to use the conditional entropy $H[y|x]$ to measure how much uncertainty raised due to observing input x. However, the entropy of a mixture random variable is hard (or intractable)) to compute because of the sum in the expression. But under mild conditions, we can show that y, as an average of the combination of weights and components, converge to a normal distribution so that the $cov(y|x)$, which is easy to compute, can be used to quantify the uncertainties in the mixture random variables. Let $\phi_1, ..\phi_K$ be the i.i.d samples that encapsulates the coefficient-component pairs where $\phi_i = \begin{pmatrix}\pi_i\mu_{1i}\\ ...\\ \pi_i\mu_{Li}\end{pmatrix}$ with mean $\mathbf{m} = \begin{pmatrix}m_1\\ ...\\ m_l\end{pmatrix} = \begin{pmatrix}\mathbb{E}[\pi_i\mu_{1i}]\\ ...\\ \mathbb{E}[\pi_i\mu_{Li}]\end{pmatrix}$ and covariance Σ. Let $y_{|x} = \begin{pmatrix}\overline{y}_1\\ ...\\ \overline{y}_L\end{pmatrix}$ where $\overline{y}_L = \sum_{k=1}^K \pi_k\mu_{kl}$. Then, according to the Multivariate Central Limit Theorem, we have $y \sim \mathbb{N}(\frac{\mathbf{m}}{K}, \frac{\Sigma}{K})$. So we can leverage the dominanant term in the entropy of multivariant Gaussian, $\ln(|[\text{cov}(y_{|x})|)$ as a surrogate measurement of $H[y_{|x}]$, since the number of components is relatively large.

The selection score is $\mathcal{A}_{\hat{\mathbf{y}}}(\mathbf{x}) = \log|\text{cov}[\hat{\mathbf{y}}|\mathbf{x}]|$. This criterion captures the expected information gain evaluated on the final label predictions when including the unlabeled samples. It aggregates the predictions from the weight coefficient predictor and the proxy pseudo-count predictor, and shapes the fine-grained uncertainty in a global view.

By focusing on epistemic uncertainty from the evidential model for weight coefficients, differences in label clusters indicated by proxy pseudo counts, and the covariance in label predictions, we devise a comprehensive multi-source uncertainty-based selection score (MSU) $\mathcal{A}(\mathbf{x}) = \mathcal{A}_{\pi_k}(\mathbf{x}) + \lambda\mathcal{A}_{\mu}(\mathbf{x}) + \eta\mathcal{A}_{\hat{\mathbf{y}}}(\mathbf{x})$. Compared to a single uncertainty score, the integration of these uncertainty measures facilitates a targeted exploration of the data space and enables the identification of the most informative samples within a constrained budget.

3.4 Algorithm Summary

We provide our proposed model and AL strategy as two main algorithms: the first one describes the entire AL process, while the second one describes the detailed training process of EMM:

Input : Total number of AL rounds: T,
Active learning budget: $B = n_b \cdot T$
Unlabeled pool: S_U,
Model at step t: $f_{\theta_t}(\mathbf{x})$,
(including encoder $e(\mathbf{x})$, decoder 1 and 2 $g_{\boldsymbol{\pi}}(\mathbf{x}), g_{\boldsymbol{\mu}}(\mathbf{x})$)
AL sampling strategy: $A : f_{\theta_t}(\mathbf{x}) \times S_U \to \mathbb{R}$,
Learning objective $\mathcal{L}$: $\mathcal{L} : f_{\theta_t}(\mathbf{x}) \times \mathbf{y} \to \mathbb{R}$,
Annotation method: $h : \mathbf{x} \to \mathbf{y}$
Output : Annotated training dataset: S_L
Randomly select S_L `// Even-label Split Method for Train/Test`
for $t = 1$ **to** T **do**
  Train preset components $\boldsymbol{\Theta}_0$ on S_L `// E-M: Bernoulli Mixture Training`
  Compute preset weights $\boldsymbol{\Pi}_0$ with $\boldsymbol{\Theta}_0$ `// Linear Program Optimization`
  ($\boldsymbol{\Theta}_0 \sim k \times l$, $\boldsymbol{\Pi}_0 \sim n \times k$, $\boldsymbol{\Pi}_0 \times \boldsymbol{\Theta}_0 = \mathbf{y}' \sim n \times l$, minimizing $\|\mathbf{y}' - \mathbf{y}\|_2$)
  Pre-train NN model ($e()$ and $g_{\boldsymbol{\pi}}()$) to fit $\boldsymbol{\Pi}_0$ `// MSE Loss of` $\boldsymbol{\Pi}_0$
  Jointly train NN model ($e()$ and $g_{\boldsymbol{\pi}}(), g_{\boldsymbol{\mu}}()$) to fit $\mathbf{y}$ `// Alternating between MSE(`$\boldsymbol{\Pi}_0$`) and Label Loss`
  (Evaluate)
  Active sample b_t from S_U based on $\mathcal{A}$
  Update the pool and training set
  $S_U = S_U \backslash b_t, S_L = S_L \cup b_t$
end

Algorithm 1: Active Learning with Evidential Multi-label Model (Outer Loop)

Input : Model: $f_{\theta_t}(\mathbf{x})$,
(including encoder $e(\mathbf{x})$, decoder 1 and 2 $g_{\boldsymbol{\pi}}(\mathbf{x}), g_{\boldsymbol{\mu}}(\mathbf{x})$)
Learning objective $\mathcal{L}$: $\mathcal{L} : f_{\theta_t}(\mathbf{x}) \times y \to \mathbb{R}$,
(Types: $\mathcal{L}_{MSE}(\boldsymbol{\Pi}_0)$, soft margin $\mathcal{L}_{MSM}(\mathbf{y})$)
Output : Model f, predictions $\boldsymbol{\Pi}_0$,$\mathbf{y}$
for $i = 1$ **to** $epochs_{pretrain}$ **do**
  Train $e()$ and $g_{\boldsymbol{\pi}}()$ using $\mathcal{L}_{MSE}(\boldsymbol{\Pi}_0)$
end
for $j = 1$ **to** $epochs_{train}$ **do**
  for $k = 1$ **to** $epochs_l$ **do**
    Freeze $e()$ and $g_{\boldsymbol{\pi}}()$
    Train $g_{\boldsymbol{\mu}}()$ using $\mathcal{L}_{MSM}(\mathbf{y})$
    ($\boldsymbol{\Theta}(\mathbf{x})$ is predicted from $g_{\boldsymbol{\mu}}()$ for each point)
    ($\boldsymbol{\Theta}_0$ is updated and saved if use book keep, otherwise do not save)
    ($\boldsymbol{\Theta}_0' = \boldsymbol{\Theta}_0 + \sum_n w_{new} \boldsymbol{\Theta}(\mathbf{x}_n)$)
  end
  for $l = 1$ **to** $epochs_\pi$ **do**
    Train $e()$ and $g_{\boldsymbol{\pi}}()$ using $\mathcal{L}_{MSE}(\boldsymbol{\Pi}_0)$
  end
end

Algorithm 2: Evidential Multi-label Model Training (Inner Loop)

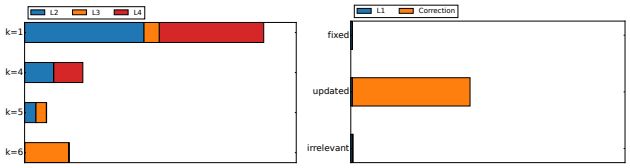
4 EXPERIMENTS

4.1 Synthetic Data

To demonstrate the effectiveness of the EMM model and the MSU sampling strategy, we design a synthetic dataset that can verify each of the proposed functionalities. The synthetic dataset contains mostly geometric feature related labels, along with a few carefully designed labels. These include a rare label **L1**, which has a frequency as low as 5% of a regular label; a couple of highly-correlated labels **L2**, **L3** which share similar features; and **L4**, where **L4**= **L2** $\cup \neg$**L3** is dependent on previous labels instead of input features.

Capturing label correlations In one example of our experiments, we train the EMM with 6 label clusters. Among these clusters, one has a particularly high weight $\mu_{\{1,L2\}}$ for **L2**. Simultaneously, the weight $\mu_{\{1,L3\}}$ for **L3** is low while the weight $\mu_{\{1,L4\}}$ for **L4** is high. Such behavior will ensure that the co-appearing **L2** and **L4** are captured during the prediction process. In the 6-cluster setting, only $\boldsymbol{\mu}_{\{1,.\}}$ and $\boldsymbol{\mu}_{\{3,.\}}$ have higher weights for **L2**, and in both clusters $\mu_{\{.,L4\}}$ is similarly high as $\mu_{\{.,L2\}}$. Meanwhile, in $\boldsymbol{\mu}_{\{6,.\}}$ we have a high weight $\mu_{\{3,L3\}}$ for **L3**, in which case $\mu_{\{3,L4\}}$ for **L4** is low (« 0.01). Furthermore, in cluster 4 $\boldsymbol{\mu}_{\{4,.\}}$, both $\mu_{\{4,L2\}}$ and $\mu_{\{4,L3\}}$ have slightly higher values but $\mu_{\{4,L4\}}$ is low (« 0.01) so **L4** does not falsely appear.

We further run a few different configurations with various numbers of components and quantify the correlations between these labels described by the label clusters. The results show that the average CosineSimilarity$(\mu_{\{.,L2\}}, \mu_{\{.,L4\}})$ is 0.91 meaning that the positive correlation between **L2** and **L4** is captured most of the times. On the other hand, the average sim$(\mu_{\{.,L3\}}, \mu_{\{.,L4\}})$ is as low as 0.12 showing the lack of correlation between **L3** and **L4**. More specifically, when neither $\mu_{\{.,L2\}}$ or $\mu_{\{.,L3\}}$ is insignificant, $R(\mu_{\{.,L2\}}, \mu_{\{.,L4\}})$ drops to 0.31, indicating that the fine relationship of "if and only if" is well-captured by the mixture model.



(a) Label cluster components concerning L2,L3,L4 **(b) Label cluster components concerning L1**

Figure 3: (a) A visualization of the labels clusters concerning L2, L3, and L4; (b) A visualization of the labels clusters concerning L1 with and without updating with proxy pseudo-counts. "fixed" is the original unsupervisedly trained $\mu_{1,L1}$, "updated" is an updated $\mu_{1,L1}(\mathbf{x}_1)$ where $\pi_1(\mathbf{x}_1) = 0.83$ meaning $\boldsymbol{\mu}_{1,.}$ is dominant, and "irrelevant" is an updated $\mu_{1,L1}(\mathbf{x}_2)$ where $\pi_1(\mathbf{x}_2) = 0.18$ meaning this cluster $\boldsymbol{\mu}_{1,.}$ does not contribute much to the prediction of $\mathbf{x}_2$.

Prediction enhancement by proxy pseudo-count combination Although the mixture model is great at capturing the label correlations with label clusters, it is not always good at predicting rare labels. For example, in the 6-cluster setting, the largest weight for **L1** is $\mu_{\{2,L1\}} = 0.016$ because of the extremely imbalanced label

Table 1: The relationship between average uncertainty scores, label cardinality, and rare labels

	$\|\mathbf{y}\|<3$	$\|\mathbf{y}\| \geq 3$	$\mathbf{y}_{L1} = 1$	$\mathbf{y}_{L1} = 0$
Average $\mathcal{A}_{\pi_k}(\mathbf{x})$	46.5	36.1	52.9	31.9
Average $\mathcal{A}_{\hat{\mathbf{y}}}(\mathbf{x})$	0.079	0.070	0.138	0.069

distribution. In this case, even if the model predicts solely $\boldsymbol{\mu}_{\{2,.\}}$ for a sample $\mathbf{x}$ ($\pi_2 = 1, \pi_k = 0, k \neq 2$), the prediction of **L1** is 0.016. This small value creates difficulty in converting the predicted score to a positive label prediction.

Rare label and correlation discovery by uncertainty quantification As for actively selecting data samples, we study the correlations between $\mathcal{A}_{\pi_k}(\mathbf{x})$, $\mathcal{A}_{\hat{\mathbf{y}}}(\mathbf{x})$, and the true labels of $\mathbf{x}$. We show a set of statistics in Table 1 to analyze these behaviors. For $\mathcal{A}_{\pi_k}(\mathbf{x})$, we compare the sampling score with the estimated unknown information of the corresponding pool samples. The unknown information is evaluated from both the feature and label perspective, using the feature similarity and the label cardinality. The correlation between the feature similarity and the uncertainty score is -0.73, and the correlation between the label similarity and the uncertainty score is -0.41. The label similarity is less indicative as the labels are $0, 1$, but we can still conclude the negative correlation between the similarity and uncertainty. For $\mathcal{A}_{\hat{\mathbf{y}}}(\mathbf{x})$, we specifically focus on the rare labels. On average, the samples containing less than 3 labels have an uncertainty score 28.8% higher than the other samples and the samples containing rare labels have an uncertainty score 65.8% higher than the regular samples.

4.2 Real Data

Datasets and experiment settings. We conduct AL experiments on representative real-world datasets including Delicious, Bibtex, Corel5k, Enron, and NUS-WIDE, covering multiple application domains [29]. The number of labels ranges from 53 to 156, most of which are relatively rare in the entire dataset. We summarize the datasets and data preprocessing in Appendix C.

Performance comparison. We first compare the AL performance with competitive multi-label AL baselines:

- **GP-B^{2}M** uses a Bayesian mixture model and conducts active sampling using the combined predicted variance of multi-output GP and the label clusters [26].
- **MMC** is model-adaptive (implemented with label ranking model) and involves a predictive process for the number of labels. It samples instances based on the expected loss [33].
- **Adaptive** uses an SVM model and considers both the SVM margin and the label cardinality inconsistency for data sampling [16].
- **CVIRS** combines the difference margin ranking (confidence) and the label vector inconsistency for data sampling [22].

For AL comparison, we use the area under the ROC curve as the main criterion. We start with 2% initially labeled samples for datasets Delicious, Bibtex, Corel5K, and 0.03% for NUS-WIDE. The initial labeled set contains a minimum of one positive instance per label to ensure that binary solutions can be trained. For EMM and methods that can perform batch active learning, we sample 5 rounds with 100 samples selected in each round. For single-batch baseline methods, we sample 500 rounds to obtain the same number of labels in the end. The base performance of classification models varies as some baseline methods use binary-SVMs (Adaptive, CVIRS), some use strategy-specific models such as the label-ranking model (MMC) and the GP-B^{2}M model.

We also include a configuration EMM-entropy that uses the proposed EMM model and a simple entropy-based selection strategy as an additional baseline, showcasing the performance gain from the proposed sampling (MSU selection) on its own.

From Figure 4, we can see that the EMM model makes better predictions using the same amount of initial labels compared to the SVM model, which explains the advantage at the starting point. Although the label ranking model or the GP-B^{2}M model may also have good performance at the starting point, they are restricted by specific sampling methods. To separately verify the advantage brought by the uncertainty quantification, we show that our MSU selection always has an advantage in selecting better AL samples compared to a simple uncertainty-based selection (EMM-Entropy).

For a more fine-grained analysis of the model performance, we also compute the average precision improvement as shown in [26]. This shows how the rare-label predictions have improved using the EMM model compared to a fully Bayesian mixture model where the label clusters are completely global.

$$\text{API}_l(\%) = \frac{\text{AP}_l(\textbf{EMM}) - \text{AP}_l(\textbf{GP-B}^2\textbf{M})}{\text{AP}_l(\textbf{GP-B}^2\textbf{M})} \times 100\% \tag{9}$$

In Figure 5, we show the API metric for the 50 rarest labels on each dataset. The improved API on a label is shown by a blue bar above the $API = 0$ axis, while the worse API performances are shown by the orange bars. The x axis shows the number of times each label has appeared in the testing set. We can see that EMM has a significant advantage on rarer labels.

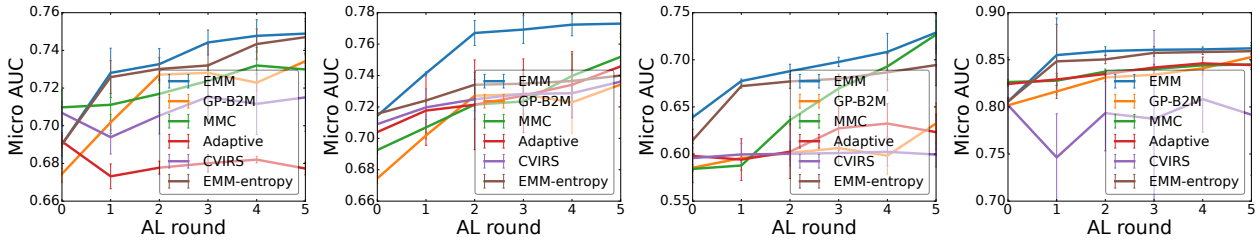
We further conduct an ablation study to demonstrate the effectiveness of the proposed methods.

Ablation Study We conduct the ablation study on model components (weight coefficient predictor and proxy pseudo-count predictor) and the AL sampling method (balancing parameters λ and η).

Since the proposed EMM model combines the evidential weight coefficient learning and the proxy pseudo-counts learning, we compare the complete model with two weakened configurations:

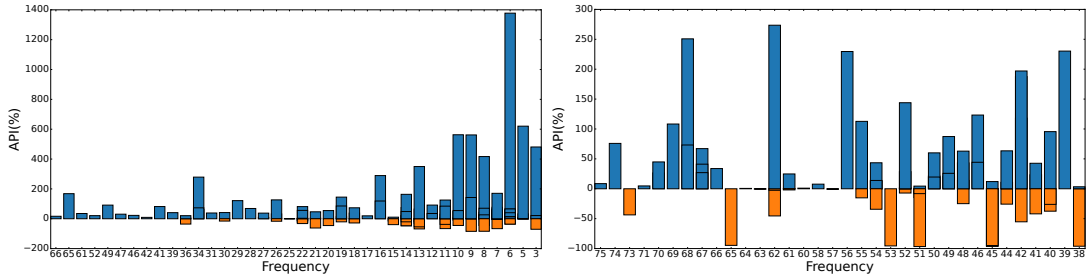
- **EMM**$^{-rr}$ reduces the weight coefficient leaner to a simple ridge regression model, and only combines the prediction with fixed Bernoulli mixtures as the label clusters.
- **EMM**$^{-fixed}$ uses the evidential learning of weight coefficients, but still only uses the fixed Bernoulli mixtures as the label clusters.

From Figure 6, we can see that the evidential regression model predicts the weight coefficients better than a simple regression model such as Ridge Regression, which shows the effectiveness of the first branch of EMM ($e()$ and $g_{\boldsymbol{\pi}}()$). We can also see that without the proxy pseudo-count updates, the performance is not as good, which shows the effectiveness of the second branch of EMM ($g_{\boldsymbol{\mu}}()$) and the joint training of the entire model.



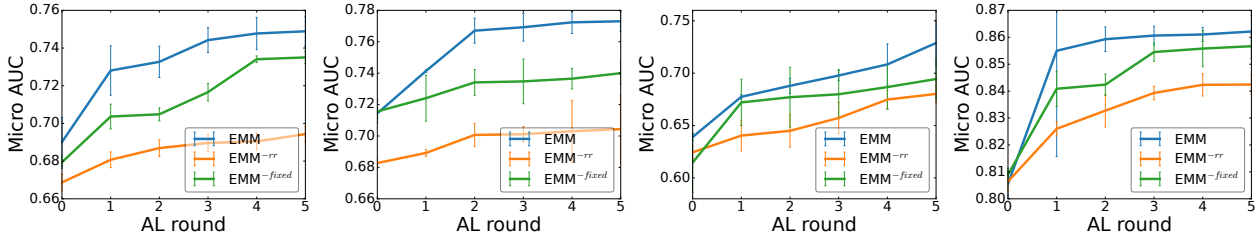
(a) Delicious (b) Corel 5k (c) BibTex (d) NUS-WIDE

Figure 4: Active learning performances on real-world datasets (AU-ROC increases as we sample 5 rounds with 100 samples in each round)



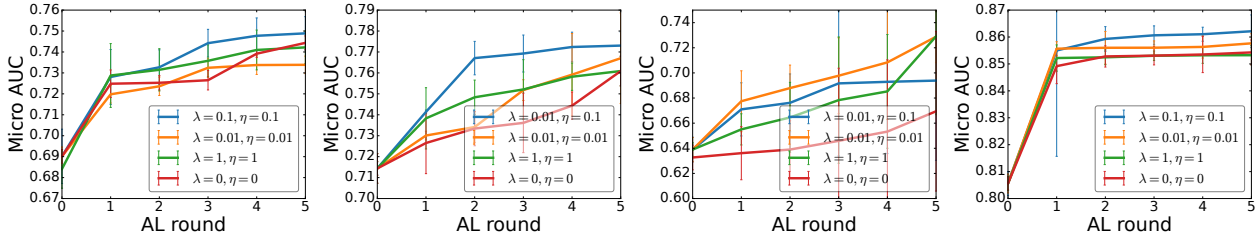
(a) Corel 5k (b) Delicious

Figure 5: Average precision improvement (API) for rare labels



(a) Delicious (b) Corel 5k (c) BibTex (d) NUS-WIDE

Figure 6: Ablation study on model components



(a) Delicious (b) Corel 5k (c) BibTex (d) NUS-WIDE

Figure 7: Ablation study on balancing parameters

From Figure 4, we can already see that the proposed evidential uncertainty-based sampling outperforms a simple metric such as Entropy. From Figure 7, we can see that our MSU sampling strategy effectively benefited from the multi-source uncertainty information compared to the single-source uncertainty ($\lambda = \eta = 0$). However, the epistemic uncertainty from the evidential model is still the most important source of uncertainty as the sampling performance decreases when we increase the balancing parameters for $\mathcal{A}_{\mu}(\mathbf{x})$ and $\mathcal{A}_{\hat{\mathbf{y}}}(\mathbf{x})$ too much.

We can also evaluate the quality of uncertainty estimation using true labels of the pool samples after AL experiments.

Real-world Uncertainty Evaluation Here, we present an example of the rare-label uncertainty being captured by both the cluster difference prediction and the covariance of the predicted

labels. From Table 2 (results obtained on BibTex), we can see that

Table 2: The relationship between average uncertainty scores, label cardinality, and rare labels

	$\|\mathbf{y}\|<3$	$\|\mathbf{y}\| \geq 3$	$\mathbf{y}_{L1} = 1$	$\mathbf{y}_{L1} = 0$
Average $\mathcal{A}_{\pi_k}(\mathbf{x})$	3.886	3.062	3.834	3.196
Average $\mathcal{A}_{\hat{\mathbf{y}}}(\mathbf{x})$	0.021	0.064	0.018	0.013

similar to the synthetic data case, the uncertainty metric captures the rare labels well, although the difference is smaller because there are more labels in total.

5 CONLUSION

In this work, we introduced a novel Evidential Mixture Machines (EMM) model, which integrates deep evidential learning within the multi-label classification framework of AL. This approach significantly enhances the handling of large and sparse label spaces, particularly addressing the challenges posed by sparse and rare labels. Our model's sophisticated uncertainty quantification and improved prediction accuracy set it apart from traditional Binary Relevance Models and other existing methodologies.

The effectiveness of the EMM model is demonstrated through rigorous evaluations on both synthetic and real-world datasets, showcasing its superiority in diverse labeling scenarios. This advancement not only contributes to the development of more efficient MLC methods but also paves the way for future research in this domain. The potential for scaling this approach to larger datasets and adapting it to various domains offers exciting opportunities for further exploration and refinement.

REFERENCES

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. 2020. Deep evidential regression. *Advances in Neural Information Processing Systems* 33 (2020), 14927–14937.

[2] Wentao Bao, Qi Yu, and Yu Kong. 2021. Evidential Deep Learning for Open Set Action Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 13349–13358.

[3] Bertrand Charpentier, Oliver Borchert, Daniel Zügner, Simon Geisler, and Stephan Günnemann. 2021. Natural Posterior Network: Deep Bayesian Uncertainty for Exponential Family Distributions. *arXiv preprint arXiv:2105.04471* (2021).

[4] Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. 2020. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in Neural Information Processing Systems* 33 (2020), 1356–1367.

[5] Junyu Chen, Shiliang Sun, and Jing Zhao. 2018. Multi-label Active Learning with Conditional Bernoulli Mixtures. In *Pacific Rim International Conference on Artificial Intelligence*. Springer, 954–967.

[6] Nengneng Gao, Sheng-Jun Huang, and Songcan Chen. 2016. Multi-label active learning by model guided distribution matching. *Frontiers of Computer Science* 10, 5 (2016), 845–855.

[7] Bin Gu and Victor S Sheng. 2016. A robust regularization path algorithm for ν-support vector classification. *IEEE Transactions on neural networks and learning systems* 28, 5 (2016), 1241–1248.

[8] Bin Gu, Victor S Sheng, and Shuo Li. 2015. Bi-parameter space partition for cost-sensitive SVM. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*. Citeseer.

[9] Bin Gu, Xingming Sun, and Victor S Sheng. 2016. Structural minimax probability machine. *IEEE Transactions on Neural Networks and Learning Systems* 28, 7 (2016), 1646–1656.

[10] Francisco Herrera, Francisco Charte, Antonio J Rivera, María J Del Jesus, Francisco Herrera, Francisco Charte, Antonio J Rivera, and María J del Jesus. 2016. *Multilabel classification*. Springer.

[11] Yibo Hu and Latifur Khan. 2021. Uncertainty-aware reliable text classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 628–636.

[12] Yibo Hu, Yuzhe Ou, Xujiang Zhao, Jin-Hee Cho, and Feng Chen. 2021. Multidimensional uncertainty-aware evidential neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 7815–7822.

[13] Sheng-Jun Huang and Zhi-Hua Zhou. 2013. Active query driven by uncertainty and diversity for incremental multi-label learning. In *2013 IEEE 13th International Conference on Data Mining*. IEEE, 1079–1084.

[14] Yeachan Kim and Bonggun Shin. 2022. In Defense of Core-set: A Density-aware Core-set Selection for Active Learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 804–812.

[15] Cheng Li, Bingyu Wang, Virgil Pavlu, and Javed Aslam. 2016. Conditional bernoulli mixtures for multi-label classification. In *International conference on machine learning*. 2482–2491.

[16] Xin Li and Yuhong Guo. 2013. Active Learning with Multi-Label SVM Classification.. In *IjCAI*. Citeseer, 1479–1485.

[17] Chunfeng Lian, Su Ruan, and Thierry Denœux. 2015. An evidential classifier based on feature selection and two-step classification strategy. *Pattern Recognition* 48, 7 (2015), 2318–2327.

[18] Nis Meinert, Jakob Gawlikowski, and Alexander Lavin. 2023. The unreasonable effectiveness of deep evidential regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 9134–9142.

[19] Deep Shankar Pandey and Qi Yu. 2022. Multidimensional Belief Quantification for Label-Efficient Meta-Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14391–14400.

[20] Deep Shankar Pandey and Qi Yu. 2023. Evidential conditional neural processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 9389–9397.

[21] Guo-Jun Qi, Xian-Sheng Hua, Yong Rui, Jinhui Tang, and Hong-Jiang Zhang. 2008. Two-dimensional active learning for image classification. In *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 1–8.

[22] Oscar Reyes, Carlos Morell, and Sebastián Ventura. 2018. Effective active learning strategy for multi-label learning. *Neurocomputing* 273 (2018), 494–508.

[23] Erik Schultheis, Marek Wydmuch, Rohit Babbar, and Krzysztof Dembczynski. 2022. On missing labels, long-tails and propensities in extreme multi-label classification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1547–1557.

[24] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems* 31 (2018).

[25] W. Shi, X. Liu, and Q. Yu. 2017. Correlation-Aware Multi-Label Active Learning for Web Service Tag Recommendation. In *2017 IEEE International Conference on Web Services (ICWS)*. 229–236.

[26] Weishi Shi, Dayou Yu, and Qi Yu. 2021. A Gaussian process-Bayesian Bernoulli mixture model for multi-label active learning. *Advances in Neural Information*

Processing Systems 34 (2021), 27542–27554.

[27] Weishi Shi and Qi Yu. 2019. Fast Direct Search in an Optimally Compressed Continuous Target Space for Efficient Multi-Label Active Learning. In *International Conference on Machine Learning*. 5769–5778.

[28] Mohan Singh, Eoin Curran, and Pádraig Cunningham. 2009. *Active learning for multi-label image annotation.* Technical Report. University College Dublin. School of Computer Science and Informatics.

[29] Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Jozef Vilcek, and Ioannis Vlahavas. 2011. Mulan: A java library for multi-label learning. *The Journal of Machine Learning Research* 12 (2011), 2411–2414.

[30] Deepak Vasisht, Andreas Damianou, Manik Varma, and Ashish Kapoor. 2014. Active learning for sparse bayesian multilabel classification. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 472–481.

[31] Dongxia Wu, Ruijia Niu, Matteo Chinazzi, Alessandro Vespignani, Yi-An Ma, and Rose Yu. 2023. Deep bayesian active learning for accelerating stochastic simulation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2559–2569.

[32] Jian Wu, Anqian Guo, Victor S Sheng, Pengpeng Zhao, Zhiming Cui, and Hua Li. 2017. Adaptive low-rank multi-label active learning for image classification. In *Proceedings of the 25th ACM international conference on Multimedia*. 1336–1344.

[33] Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen. 2009. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 917–926.

[34] Min-Ling Zhang and Zhi-Hua Zhou. 2006. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE transactions on Knowledge and Data Engineering* 18, 10 (2006), 1338–1351.

[35] Yi Zhang. 2010. Multi-task active learning with output constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 24. 667–672.

[36] Yu Zhang, Bowen Jin, Xiusi Chen, Yanzhen Shen, Yunyi Zhang, Yu Meng, and Jiawei Han. 2023. Weakly Supervised Multi-Label Classification of Full-Text Scientific Papers *(KDD '23)*. Association for Computing Machinery, New York, NY, USA, 3458–3469. https://doi.org/10.1145/3580305.3599544

Appendix

A APPENDIX ORGANIZATION

Organization. In this appendix, we provide additional details of the work, including a summary of notations in Appendix B and the experiment details with additional results in Appendix C.

B SUMMARY OF NOTATIONS

Table 3: Summary of key notations with definitions

Notation	Definition
$\mathbf{y} = (y_1, ..., y_L)^\top$	Multi-label vector
$\mathbf{x}$	Data feature vector
$\mathbf{z}_k = \prod_{l=1}^{L} \text{Bernoulli}(y_l, \mu_{kl})$	L-variate Bernoulli random variable.
μ_{kl}	Beta random variable
(a_{kl}, b_{kl})	Parameters of the Beta distribution.
π_k	Gaussian random variable (component coefficient).
$\mathbf{p} = (\gamma, \nu, \alpha, \beta)$	Parameters of inverse normal gamma distribution.
$\mathcal{N}(\tau\|\gamma, \frac{\sigma^2}{v})$	Gaussian distribution in the evidential NIG prior with mean γ and variance of $\frac{\sigma^2}{v}$
$\Gamma^{-1}(\sigma^2\|\alpha, \beta)$	Inverse Gamma distribution in the evidential NIG prior with shape parameter α and scale parameter of β
$St(\pi; \cdot, \cdot, \cdot)$	Student t distribution
Evidence($\mathcal{E}$)	The integrated model evidence
$\mathcal{A}(\mathbf{x})$	Active sampling score function

C ADDITIONAL EXPERIMENT DETAILS AND RESULTS

C.1 Experiment settings

Our experiments are performed on clusters with NVIDIA A6000 and NVIDIA A100 graphic cards and Intel Xeon Gold 6150 CPU processors. The runtime of the experiments varies depending on the size of the unlabeled pool. Compared to traditional models, the evidential deep learning model takes a longer time to train. However, compared to Bayesian models or label ranking models, the inference time is significantly shorter.

For our main results, we pre-train the label clusters using an E-M algorithm and obtain initial optimal weights for the labeled training set using linear programming optimization. The evidential model is trained for around 5000 epochs to fit the weight coefficients $\boldsymbol{\pi}$, and the joint training step is trained iteratively for 100 epochs in each round.

C.2 Synthetic Data Settings

We design the synthetic data to demonstrate the model behavior regarding rare labels and label correlations. To achieve this, we create the following set of labels: a set of geometry information-based labels (which have strong feature-label connections), a set of non-geometry information-based labels (which prevent the problem from being purely goemetry-based), and labels L_1 to L_4 which are introduced in the main paper and our main concern. As shown in Figure 8, labels L_1 to L_4 are rare and correlated.

C.3 Additional Baseline Comparison

Here we show additional AL comparison with two other baselines:

- **AUDI** uses a label ranking mechanism, where a dummy label is used to separate the positive and negative labels. Its sampling function is based on a modified cardinality inconsistency measure [13].
- **CS** uses a compressed sensing mechanism combined with GP predictions [27].

From Figure 9 we can see that both baselines do not perform well at the low budget, while EMM still outperforms them. The implementation of AUDI and CS on BibTex and NUS-WIDE are computationally too demanding and we would expect similar behavior given the labeling budget.

We also present the API results on more datasets:

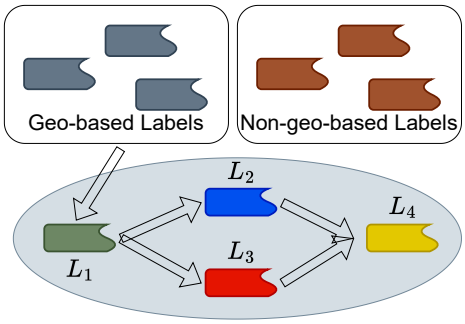
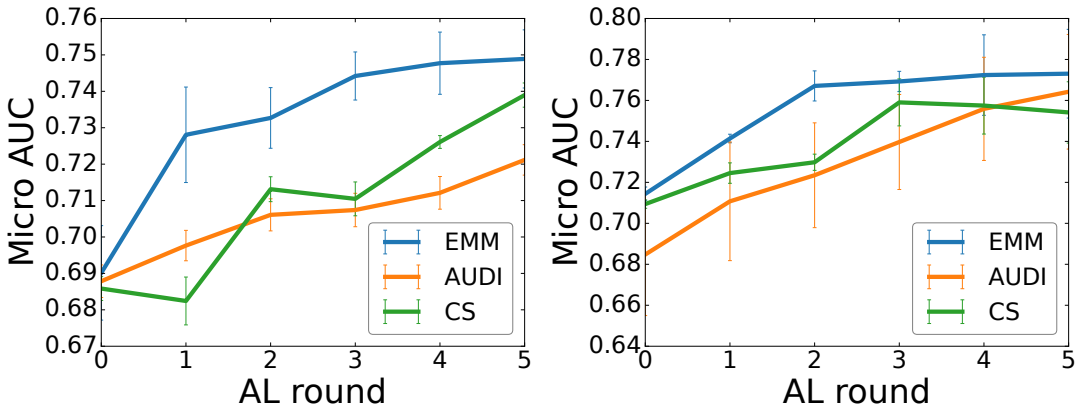
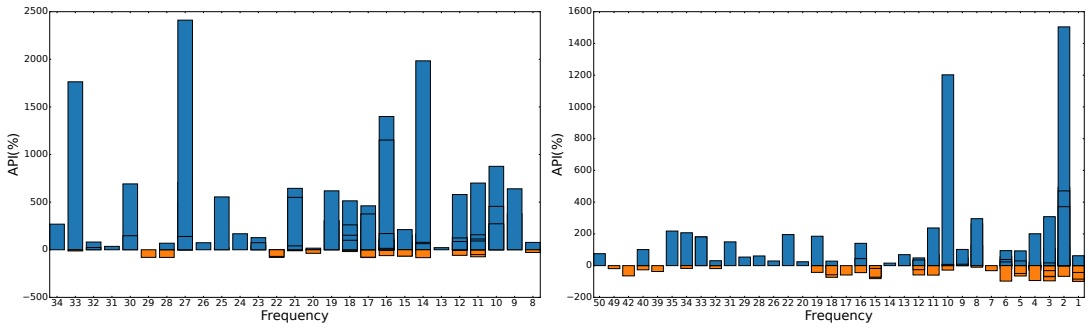


Figure 8: A visualization of label composition in the synthetic dataset



(a) Delicious **(b) Corel 5k**

Figure 9: Additional active learning performances comparison on smaller0sized real-world datasets (the AUDI and CS baselines become computationally expensive on larger datasets)



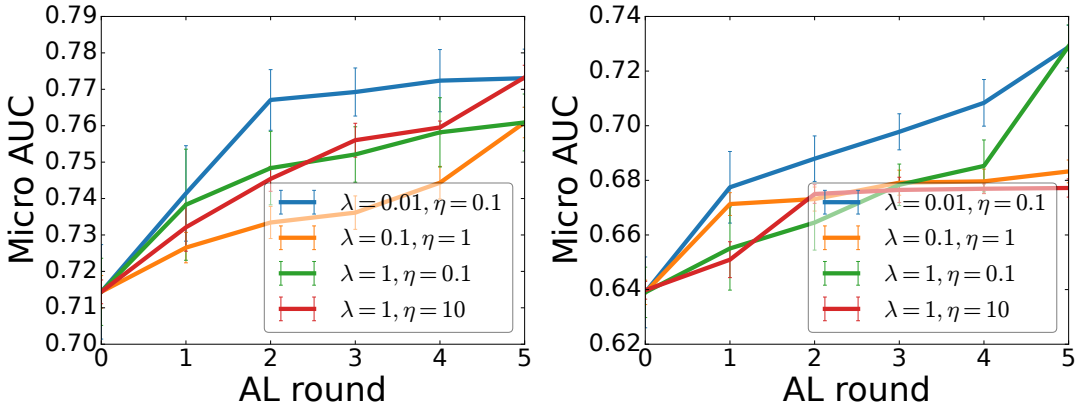
(a) BibTex **(b) NUS-WIDE**

Figure 10: Average precision improvement (API) for rare labels

C.4 Additional Ablation Study

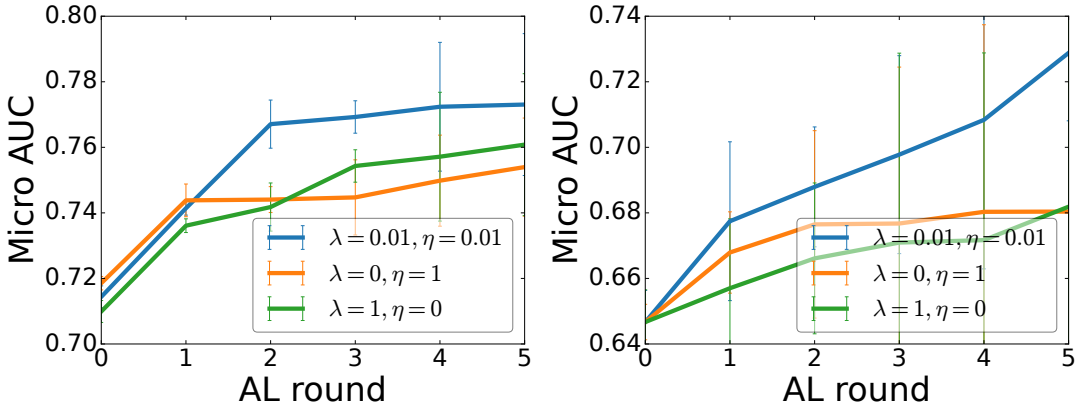
Here we show results using more configurations of λ and η. As explained in the main paper, these balancing parameters work well with a moderately small value. With λ and η both around 0.1 to 0.01, we are able to obtain stable AL results. However, if the values are set too large, the performance may degrade.

When we set $\lambda = 0, \eta = 1$ or $\lambda = 1, \eta = 0$, we get the combination of $\mathcal{A}_{\pi_k}$ and a single source of label uncertainty $\mathcal{A}_{\mu}$ or $\mathcal{A}_{\hat{y}}$. As we can see, the combination of all three works the best.



(a) Corel 5k **(b) BibTex**

Figure 11: Additional ablation study on balancing parameters: different values



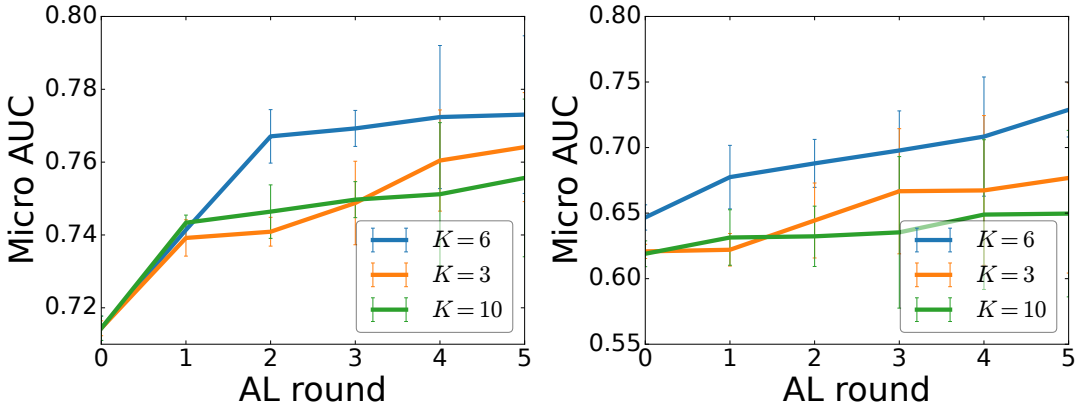
(a) Corel 5k **(b) BibTex**

Figure 12: Additional ablation study on balancing parameters: single uncertainty source

In Figure 13, we show the comparison between different K values. As we can see, extremely low number of clusters is not sufficient for achieving good model performance, while a large number is much more costly and also suffers from overfitting. The latter problem might harm the AL sampling more as we see the $K = 10$ case performs even worse than $K = 3$.

D SOURCE CODE

https://anonymous.4open.science/r/EMM-2367/



(a) Corel 5k **(b) BibTex**

Figure 13: Additional ablation study on number of clusters K