Interpretable Sample Selection with Exceptional Model Mining

Vincent P. Hoogendam[1], Kalina R. Bakardzhieva[1], Luca Mainardi[1], Luyang Xie[1], and Rianne M. Schouten[1](✉)

Eindhoven University of Technology, the Netherlands
{v.p.hoogendam,k.bakardzhieva,l.mainardi,l.xie}@student.tue.nl;
r.m.schouten@tue.nl

Abstract. Active Learning (AL) aims to improve model performance while minimizing labeling effort by selecting informative samples into the training dataset. The importance is particularly clear when humans are involved in the data collection and labeling process, and even more so in the context of human-computer interaction (HCI) where users can provide only few training samples and those samples should be as useful as possible. Traditionally, AL methods iteratively increase the training data set by selecting informative samples from an unlabeled dataset. However, AL methods may be prone to oversampling certain regions of the data space, potentially introducing redundancy in the dataset. To address this issue, we propose an interpretable sample selection approach based on a local pattern mining framework called Exceptional Model Mining (EMM). EMM aims to discover subgroups in the dataset that somehow behave exceptionally. These subgroups are described using an interpretable language of conjunctions of attribute-value pairs. We propose an EMM-based AL approach that discovers, describes and selects informative samples. As such, our method provides an explanation as to why model performance is reduced for certain samples. In addition, our method is an alternative to existing AL methods: given a diverse subgroup set, we create a diverse and representative training set by selecting samples from each subgroup. We evaluate the performance of our approach against a traditional AL baseline and demonstrate that it provides interpretable explanations and has good classification power, especially when labeling budget is low.

Keywords: Active Learning · Exceptional Model Mining · Interpretable Machine Learning · Sample Selection · Explainable AI

1 Introduction

The rapid expansion of machine learning (ML) is transforming various industries, driving advancements in areas such as healthcare, finance, and autonomous systems [6, 1, 21, 17]. Much of the success is attributed to developments in supervised learning, where a predictive model $M : \mathcal{X} \to \mathcal{Y}$ is learned based on a training dataset $D_{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ of n IID samples. The quality of D_{train}

is key for the quality of M. However, obtaining a sufficient number of informative samples could be time consuming, expensive, and require large amounts of manual, human effort or expertise. This problem is particularly apparent when humans are involved in the data collection and labeling process, and even more so in the context of human-computer interaction (HCI), for instance when a machine must learn from a small number a training samples [7, 26, 3].

To improve the quality of training data D_{train}, Active Learning (AL) aims to iteratively select samples that are most informative [33, 36]. Typically, the model M_{train} trained on an initial dataset D_{train} is evaluated on a pool of unlabeled samples $D_{\text{unlabeled}}$, from which a small set of informative samples are selected and labeled [22, 23, 19]. The model M_{train} is then updated (or, re-trained) using these selected samples, ideally resulting in an improved performance. The idea is that when a model is presented with a training sample that is very similar to the ones it has already seen, it is unlikely to learn from it; in contrast, by actively selecting which samples to label, AL reduces the total number of samples needed while simultaneously improving the model's quality. In principle, the process of selecting samples and updating the model can be repeated indefinitely.

AL faces challenges that limit its efficiency. Traditional methods based on uncertainty sampling (i.e., selecting samples of which the model is least certain) [5, 19, 4] and Query-By-Committee (i.e., constructing a committee of models and selecting samples where the committee disagrees about the prediction) [35] could lead to redundant sample selection, increased computational costs, and bias in sample selection. These limitations can result in inefficient use of the labeling budget without providing enough informative samples for learning. Furthermore, existing AL methods offer little interpretability as to why a certain sample is selected [28, 24, 13].

Therefore, in this paper, we propose an interpretable sample selection method based on a local pattern mining framework called Exceptional Model Mining (EMM) [9, 31]. EMM aims to discover subgroups in a dataset that somehow behave exceptionally. These subgroups are described using interpretable, rule-based conjunctions of attribute-value pairs. Exceptional behavior is defined using a quality or exceptionality measure over parameters in a target model, estimated on a set of target variables.[1]

We propose to utilize the EMM framework for discovering subgroups of samples with reduced model performance. To be specific, we build our target model on the confidence scores of the unlabeled samples and propose three quality measures for selecting the samples for which the model is least confident. Our method contains two steps and our contributions are twofold:

1. We provide a diverse list of interpretable subgroups of informative samples. That is, our method not just discovers subgroups of informative samples, but additionally describes these subgroups in a human-interpretable way. We

[1] The term *target* should not be confused with outcome variable Y in a supervised learning context. In EMM, the goal (i.e., target) of our task is to detect exceptionality, but there is no ground-truth as to where in the feature space exceptional behavior may occur.

thus explain why existing AL methods select certain samples and provide insights into the characteristics of the unlabeled data space.

2. In addition, our method is an EMM-based AL approach that can be used as an alternative to existing AL methods. That is, we select samples from each subgroup and iteratively add those samples to the training dataset. Since our subgroup set is diverse, our approach contributes to solving the existing redundancy problem in AL.

The rest of this paper is organized as follows. First, we introduce notation and provide background information on EMM in Section 2. We then position our approach with respect to existing methods in EMM and AL in Section 3. In Sections 4.1 and 4.2, we detail our proposed EMM-based AL approach in two steps. We then evaluate the performance of our approach in Sections 5 and 6. Section 7 concludes.

2 Background

2.1 Preliminaries

Let $D_{\text{train}} = L = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_L}$ be the initial, labeled set, and $D_{\text{unlabeled}} = U = \{\mathbf{x}_i\}_{i=1}^{n_U}$ be the pool of unlabeled samples. Here, $\mathbf{x}_i \in \mathbb{R}^d$ are the feature vectors in a predictive model $M : \mathbb{R}^d \to \mathcal{Y}$. The features are also to used as descriptive attributes $a_1, a_2, ..., a_d$ to create and describe subgroups in the Exceptional Model Mining (EMM) [9] context, see more in Section 2.2. In this paper, we assume $\mathcal{Y} \in [0, 1]$.

We refer to M_0 as the model trained on the initial, labeled dataset L. For any unlabeled sample $\mathbf{x}_i \in U$, we then obtain a prediction $\hat{y}_i$ and a confidence score c_i where $\hat{y}_i = M_0(\mathbf{x}_i)$ and $c_i = \arg\max_{y \in \mathcal{Y}} P_{M_0}(y \mid \mathbf{x}_i)$. Here, $P_{M_0}(y \mid \mathbf{x}_i)$ is the predicted probability of class $y \in \mathcal{Y}$.

The AL task is to select n^t samples from U such that the model can be updated from M_{t-1} to M_t, where t indicates the iteration. We denote the selected samples at iteration t with $V_t = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n^t}$. In this paper, we take an approach where the model is re-trained using $D_t = D_{t-1} \cup V_t$. Remark that when applying AL in the real-world, it requires effort to obtain the true labels y_i of the selected $i \in \{1, 2, ..., n^t\}$ samples in V_t. As an alternative in this paper, we use an experimental setup where L (D_{train}) and U ($D_{\text{unlabeled}}$) are obtained by randomly splitting an initial dataset D into three sets: $D_{\text{train}}, D_{\text{unlabeled}}$ and a test dataset D_{test}. We thus have the true labels of the samples in U available, and can use them to re-train the model in the next iteration.

2.2 More on Exceptional Model Mining

Exceptional Model Mining (EMM) [9, 31] is a local pattern mining framework seeking coherent subgroups in the dataset that behave exceptionally. Traditionally, EMM assumes a dataset Ω to be a bag of n records $r \in \Omega$ of the form

$$r = \Big(a_1, \ldots, a_k, \ell_1, \ldots, \ell_m\Big), \tag{1}$$

where k and m are positive integers. In EMM, we call $a_1, \ldots, a_k$ the *descriptive attributes* or *descriptors* of r, and $\ell_1, \ldots, \ell_m$ the *target attributes* or *targets* of r. In this paper, we use the predictors as descriptors; $X_1 = a_1, X_2 = a_2, ..., X_d = a_k$ and thus, $d = k$.

Subgroups are defined using descriptions; a Boolean function $S : \mathcal{A} \to \{0, 1\}$. A description S covers a record r^i if and only if $S(a_1^i, ..., a_k^i) = 1$.

Definition 1 (Subgroup cf. [9]). *A subgroup corresponding to description S is the bag of records $G_S \subseteq \Omega$ that S covers:*

$$G_S = \{r^i \in \Omega \mid S(a_1^i, a_2^i, \ldots, a_k^i) = 1\}.$$

The complement contains all non-covered records: $G_S^C = \Omega \setminus G_S$ [9, p.53]. In this paper, we discover our subgroups in the unlabeled dataset and thus, $\Omega = U$.

In EMM, the choice of description language $\mathcal{S}$ is free, though generally we let the description be a conjunction of selection conditions over the descriptors, where condition sel_j is a restriction on the domain $\mathcal{A}_j$ of the attribute a_j. For discrete variables the selector may be an attribute-value pair $(a_j = v)$; for continuous variables it could be a range of values $(w_1 \leq a_j \leq w_2)$ [9].

We aim to discover the descriptions for which the subgroups display exceptional behavior on a target model, fitted to a set of target attributes. A quality measure quantifies the exceptionality of within-subgroup behaviour with respect to some reference behaviour model.

Definition 2 (Quality Measure cf. [9]). *A quality measure (QM) is a function $\varphi : \mathcal{S} \to \mathbb{R}$ that assigns a numerical value to a description S.*

In this paper, we use the confidence scores as our target attribute, and hence $m = 1$. An example of an existing quality measure for 1 numerical target attribute is proposed by [18] and called a *t-score*: it evaluates the distance between the mean estimate in the subgroup and the mean estimate in the overall population (the entire dataset). We present our proposed quality measures in Section 4.1.

Overall, the challenge in EMM is to effectively search through the descriptive space to find the top-q best-scoring subgroups (the subgroup set). In this paper, we use the commonly used beam search algorithm cf Algorithm 1 in [9, 31]. In essence, beam search takes parameters w and d_{max}, which define the width and depth of the algorithm. It is a heuristic search algorithm that explores the search space by maintaining a fixed number of the best candidate subgroups at each level, defined by the beam width w. The algorithm works as follows: it begins with the most general subgroup and iteratively expands it by adding new conditions to create more specific subgroups, up to a maximum depth d_{max}. Then at each iteration, the algorithm first refines the subgroup by adding one feature constraint and then evaluates the exceptionality by calculating the quality value following φ. The top w subgroups will be expanded in the next iteration. By focusing on the most promising subgroups, beam search allows us to efficiently explore the search space without considering every possible option, resulting in high-quality subgroups found in reasonable time [9]. Given the advantages of beam search, the EMM challenge shifts from developing an effective search algorithm to developing the most suitable quality measures that lead the search towards the interesting, exceptional subgroups. In Section 4.1, we discuss our proposed quality measures for using EMM as an AL method. In Section 4.2 we then outline our choice for search parameter settings and pruning techniques to increase the diversity of the discovered subgroup set.

3 Related work

Active Learning (AL) is a machine learning approach used when obtaining labels for the training data is expensive. This can be due to variety of reasons, for example, if an opinion of expert is required which requires manual investigating of each sample. Another reason could be that a difficult or risky procedure is needed, for example, performing a biopsy on a patient. The main idea behind AL is to reduce the number of labeled samples needed by selecting the most informative samples to be labeled [33].

To be more precise about definitions, [11] distinguish AL, where the model *self-evaluates* the unlabeled samples, mostly based on uncertainty sampling [5, 19, 4], from the domain of Machine Teaching (MT) [40], where there exists a separate oracle that provides selected samples to the predictive model. Our proposed EMM-based AL method can be seen as as such a separate oracle. Furthermore, [11] distinguish MT from Active Teaching (AT). In AT, the oracle is an interactive MT method that iteratively learns and improves the decision-making about which samples to select [1, 20]. In our proposed approach, the subgroup search is repeated in every iteration, but there is no additional knowledge available to the search algorithm to improve the quality measures or otherwise improve the search for informative samples. Therefore, we consider our approach to be AL or MT rather than AT.

AL has been applied in various domains such as natural language processing [37, 19] and bio-informatics [6]. Others have developed AL techniques for time-varying data, such as Hidden Markov Models (HMM) [30] and sequential data [34]. However, it suffers from some problems. One of them is the cold start problem, that is, the initial model is trained on a very small datasets which can cause unreliable confidence scores. Since these are used for the selection of future samples for labeling this might cause further problems [27]. Another problem is the model bias. The active learning may over sample certain regions where the uncertainty is high and thus create bias in the model [25]. Finally, active learning could be computationally expensive since it requires retraining and using the model on the whole dataset at each iteration. In this paper, we aim to reduce bias in the model by selecting diverse samples from different regions of the unlabeled data space.

Our method builds on the framework of Exceptional Model Mining (EMM) [9], a local pattern mining approach that aims to discover subgroups in a population that somehow behave exceptionally. Generally, we consider EMM to a generalization of the task of Subgroup Discovery (SD) [12, 39, 16], which focuses on 1 target attribute where EMM assumes $m \geq 2$ attributes. The task of Subgroup Discovery (SD) was well defined as follows: "In subgroup discovery, we assume we are given a so-called population of individuals (objects, customers, ...) and a property of those individuals we are interested in. The task of subgroup discovery is then to discover the subgroups of the population that are statistically most interesting, are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest" [39, p.84-85]. In EMM, the property of interest can be defined more flexibly, but in essence the goal is still to discover interesting subgroups in a population. Since our proposed quality measures in Section 4 use the confidence scores obtained with a predictive model M, we essentially use 1 target attribute and our method could therefore be considered a specific subtype of EMM.

One of our proposed quality measures follows [32], who discover exceptional trend behavior in Repeated Cross-Sectional data. Specifically, [32] define exceptional behaviour within measurement waves using a *z-score*, where the distance between a parameter estimate in the subgroup and the population is divided by the standard error of the subgroup's estimate. In this way, [32] account not only for concept drift in the data-generating distributions of the descriptors, but additionally push the search to larger subgroups. Earlier, [29] proposed a quality measure which was called a z-score, but there the distance is divided by the standard deviation of the entire target attribute, rather than the standard deviation of the subgroup's estimate (i.e., the standard error).

Our proposed method contributes to the domain of eXplainable AI (XAI) [2] since we provide interpretable descriptions of informative samples. This information is directly valuable for improving the understanding of the trained model, as it reveals which parts of the feature space are important for selecting informative samples. Other EMM- and SD-based methods for XAI have been proposed [10, 8, 14, 15].

4 Our proposed approach

AL aims to select the most informative unlabeled samples for labeling [33]. Traditional methods often rely on measures of prediction confidence and uncertainty. In this paper, we propose using EMM to identify subgroups in the unlabeled dataset where the model performs exceptionally bad. Our method contains two steps, detailed in Sections 4.1 and 4.2 respectively. First, our method discovers a variety of interpretable subgroups and therefore provides an explanation as to why model performance is reduced for certain samples. Second, our method creates a diverse and representative training set by selecting samples from each subgroup in a diverse subgroup set.

4.1 Discovering and describing subgroups of informative samples

We aim to discover subgroups of informative samples in the unlabeled dataset U by defining a target attribute that represents the *uncertainty* of the unlabeled samples' prediction. To be specific, given the vector of confidence scores $\mathbf{c} = (c_1, c_2, ..., c_{n^U})$ of the n^U unlabeled samples in U, the target value of sample $i \in \{1, 2, ..., n^U\}$ as $\ell_i = 1 - 2 \cdot |c_i - 0.5|$. Consequently, the lower the confidence of the prediction, the higher the target values. Remark that this particular mapping from $\mathbf{c}$ to ℓ is quite specific for our assumption that $\mathcal{Y} = [0, 1]$ (see Section 2). However, the principle of changing confidence scores into uncertainty scores is an easy step that one can apply in other contexts such as a different outcome space or other decision boundaries.

We define three quality measures to quantify how exceptional a candidate subgroup S is compared to the global data. First, we propose the *mean uncertainty deviation* as follows:

$$\varphi_{\text{mud}}(S) = \mu_S - \mu_U = \frac{\sum_{i \in S} \ell_i}{|S|} - \frac{\sum_{i \in U} \ell_i}{|U|}. \tag{2}$$

Here, $|S|$ and U are the size (i.e., number of samples) of the candidate subgroup and unlabeled pool respectively. The quality measure calculates the mean deviation of the subgroup's uncertainty scores from the overall uncertainty mean of U. The smaller the confidence, the higher the uncertainty, the larger the distance between μ_S and μ_U, and the higher the exceptionality score φ_{mud}.

It is likely that a quality measure like φ_{mud} results in small subgroups, since an extreme deviation is more easily obtained by selecting few samples. Therefore, we follow [9] and propose to incorporate the entropy function φ_{ef}:

$$\varphi_{\text{ef}}(S) = -\left(\frac{|S|}{|U|} \log_2 \frac{|S|}{|U|} + \frac{|S^C|}{|U|} \log_2 \frac{|S^C|}{|U|} \right). \tag{3}$$

Here, S^C represents the complement of the subgroup. Then, the *mean uncertainty deviation with entropy correction* is:

$$\varphi_{\text{mudef}}(S) = (\mu_S - \mu_U) \cdot \varphi_{\text{ef}}(S). \tag{4}$$

Quality measure φ_{mudef} is expected to favor subgroups that not only deviate from the uncertainty norm but additionally have sufficient size.

An alternative way to push the search away from tiny subgroups, is by considering μ_S as a statistical estimate with $se(\mu_S)$ its standard error. We then create the *uncertainty z-score*:

$$\varphi_{\text{uz}}(S) = \frac{|\mu_S - \mu_U|}{\sigma_S / \sqrt{|S|}}, \tag{5}$$

where σ_S represents the standard deviation of uncertainty values of the covered samples in the subgroup. We expect φ_{uz} to be ideal for detecting subgroups of outliers or anomalous subgroup behaviors, since this approach considers the variability within the subgroup itself.

We can use an existing search strategy as discussed in Section 2.2 to discover and describe a top-q list of subgroups of samples with high uncertainty (low confidence) scores. Traditional AL methods would simply select these samples in the background. Instead, our pattern mining approach provides an explanation as to why the selected samples are selected.

4.2 An EMM-based AL method

Given the top-q list of subgroups of informative samples, we propose Algorithm 1 when using EMM as a selection mechanism in AL. In essence, the algorithm performs beam search with one of the quality measures $\varphi \in \{\varphi_{\text{mud}}, \varphi_{\text{mudef}}, \varphi_{\text{uz}}\}$ as described above in Section 4.1 (line 6), and then selects n^t samples from the list of discovered subgroups $[S_1, S_2, ..., S_q]$ (line 7). The selected samples are added to the training dataset (line 8) and removed from the unlabeled dataset (line 9). The process can continue for as many iterations as desired, until the labeling budget is exhausted or a sufficient accuracy is achieved (line 11).

There are several ways to select the n^t samples from the list of discovered subgroups. In this paper, we evaluate two strategies in Section 6.2: the first is to select samples at random from each group; the second strategy is to select the most uncertain samples within each group. Furthermore, one can vary the number of selected samples per subgroup. In Sections 6.2 and 6.3, we evaluate results for selecting 1, 2 or 5 samples per subgroup per iteration.

5 Experimental setup

The experiments utilize the Pima Indians Diabetes (PID) database, which consists of medical diagnostic data for predicting diabetes and therefore reflects a

Algorithm 1 EMM-Based Active Learning

```
Require: Labeled training set $D_0$, unlabeled dataset $U_0$, test set $T$
  1: Initiate $t = 0$
  2: $M_t \leftarrow$ Train model on $D_t$
  3: $\text{Accuracy}_t \leftarrow \text{Evaluate}(M_t, T)$
  for $t \in \{1, 2, ...\}$ do
      4: $\mathbf{c} = \{c_i = \text{Confidence of } M_{t-1} \text{ on } U_{t-1} \mid \forall\, \mathbf{x}_i \in U_{t-1}\}$
      5: $\ell = f(\mathbf{c})$
      6: $[S_1, S_2, ..., S_q] = \text{BeamSearch}(\varphi, U_{t-1} \cup \ell, d_{\max}, w, q)$
      7: $V_t = \{\text{SelectSamples}(S_j, n^t/q)\}_{j=1}^{q}$
      8: $D_t = D_{t-1} \cup V_t$
      9: $U_t = U_{t-1} \setminus Z_t$
      10: $M_t \leftarrow$ Train model on $D_t$
      11: $\text{Accuracy}_t \leftarrow \text{Evaluate}(M_t, T)$
      12: Repeat until labeling budget exhausted or sufficient Accuracy achieved
  end for
```

scenario where data is collected from real humans and labeling requires a possibly long and possibly intense diagnostic process. The data is available publicly through Kaggle.[2] All code used for running the experiments can be found on our github repository: `https://github.com/luca-mainardi/EMMAL`.

The PID dataset contains records for $n = 768$ female patients. The attributes included in the dataset capture a variety of physiological measurements relevant to diabetes diagnosis. The attributes are: the number of pregnancies each patient has had, the plasma glucose concentration (measured two hours after an oral glucose tolerance test), the diastolic blood pressure (millimeters of mercury), the triceps skin fold thickness (millimeters), the serum insulin levels (microunits per milliliter), the Body Mass Index (BMI), a diabetes pedigree function, which estimates the patient's hereditary risk of diabetes and the patient's age. All these features are used as predictors in our classification model M, and as descriptors in our EMM-based AL approach. The classification outcome variable is a binary outcome indicating the presence (1) or absence (0) of diabetes.

To create our datasets L, U and T, we randomly pick 20% of the full dataset for testing. Then out of the remaining 615 samples, only 123 or 20% are initially labeled and assigned to L. The last 492 samples are considered unlabeled in U. The classification model we choose for the experiment is Naive Bayes since it is a popular, simple and fast classification model. The accuracy on the withheld test set is recorded and proceed with our AL process following Algorithm 1. We use 10 iterations.

We first evaluate our three proposed quality measures from Section 4.1. We want to achieve group descriptions which are both interpretable and expressive so we choose to have a bin depth of $d_{\max} = 3$ for the beam search. This means each subgroup will be described by at most 3 attributes. Furthermore, we set the bin width to $w = 10$, meaning the algorithm brings 10 subgroups to the next search level. We set $q = 10$.

Furthermore, for the use of the discovered subgroups in selecting active learning samples, it is preferable that the subgroups have as little overlap as possible. In that way, the subgroups cover various parts of the feature space. Therefore, in this experiment we we compute and maximize the coverage of the subgroup list when selecting the beam and final subgroup list.

As shortly discussed in Section 4.2, we evaluate our proposed method for selecting samples from the subgroups *randomly* and *based on uncertainty value*. We furthermore perform the experiment with different number of selected samples per iteration. Given that we set $q = 10$, selecting 1 per subgroup means selecting $n^t = 10$ per iteration, selecting 2 per subgroup is equivalent to $n^t = 20$ and selecting 5 samples equals $n^t = 50$.

In Section 6.3, we compare our proposed EMM-based AL method against two baselines. First, we use a random selection mechanism (no AL) where we simply randomly select and label the required number of n^t samples from the unlabeled set. Second, we use an AL baseline where we utilize uncertainty sampling. This means that after calculating a confidence score c_i for each sample $\mathbf{x}_i \in U$, we

[2] `https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database`

identify and select the required number of instances that have a confidence score closest to 0.5. Similarly, in this paper we take the n^t samples that have the highest uncertainty ℓ_i (as described in Section 4.1).

6 Experimental results

6.1 Evaluating our three proposed quality measures

Using different quality measures in the EMM beam search results in different list of exceptional subgroups. Here, we perform a qualitative analysis to investigate how the quality measure impacts the resulting groups. The quality measures compared are mean uncertainty deviation φ_{mud} in Equation (2), mean uncertainty deviation with subgroup size correction using the entropy function φ_{mudef} as in Equation (4) and the uncertainty z-score φ_{uz} as in Equation (5). We present the top-3 for the three quality measures in Tables 1, 2 and 3 respectively.

First of all, we notice that the different quality measures produce groups of different sizes. The subgroups produced by φ_{mud} are of similar size, usually between 30 and 50 samples. Furthermore, with φ_{mud}, all of the groups are described by two attributes. The combined size of all $q = 10$ groups is 425, which indicates that there are overlapping samples regardless of the used cover-based pruning technique.

As expected, the groups produced by φ_{mudef} are significantly larger than the groups produced by φ_{mud}. The top-3 subgroups discovered with φ_{mudef} cover about 150 samples per subgroup, and the total size of all samples is 1330. Here, interestingly, all subgroups are described by a single attribute.

Similar like the groups produced by φ_{mudef}, the groups found with φ_{uz} are also larger than the ones discovered by φ_{mud} (where no subgroup size correction was applied). In contrast to φ_{mudef}, quality measure φ_{uz} leads to a subgroup list that contains more heterogeneous subgroups, that is, there is a larger variety of subgroup sizes; some being about 150 and others closer to 50. At the same time, the combined size of samples is 1045, which is larger than the combined sizes of the other quality measures, indicating the presence of some very large subgroups at a lower rank in the top-10.

Regarding the features in the descriptions of the subgroups, all quality measures rely on the variables *insulin* and *glucose*. Clearly, these variables are important for correctly predicting whether a patient (sample) has diabetes or not.

Table 1: Description of top-3 subgroups using Mean Uncertainty Deviation φ_{mud} as in Equation (2). The top-10 subgroups together cover $n = 425$ cases.

SG	Description	Score	Size
1	$130 <$ Insulin ≤ 744 AND $115 <$ Glucose ≤ 140	0.26	49
2	$18.57 <$ Glucose ≤ 140 AND $3 <$ Pregnancies ≤ 6	0.23	37
3	$130 <$ Insulin ≤ 744 AND $0 <$ Pregnancies ≤ 1	0.25	31

Table 2: Description of top-3 subgroups using Mean Uncertainty Deviation with subgroup size correction with the Entropy function φ_{mudef} as in Equation (4). The top-10 subgroups together cover $n = 1330$ cases.

SG	**Description**	**Score**	**Size**
1	$130 <$ Insulin ≤ 744	0.15	147
2	$18.57 <$ Glucose ≤ 140	0.12	153
3	$41 <$ Age ≤ 72	0.08.5	136

Table 3: Description of top-3 subgroups using Uncertainty Z-score φ_{uz} as in Equation (5). The top-10 subgroups together cover $n = 1045$ cases.

SG	**Description**	**Score**	**Size**
1	$130 <$ Insulin ≤ 744 AND $115 <$ Glucose ≤ 140	8.5.56	49
2	$130 <$ Insulin ≤ 744	8.74	147
3	$36 <$ BMI ≤ 67 AND $18.57 <$ Glucose ≤ 140	6.55	52

Other important features are *age* and *BMI*. Remark that currently, not all descriptions in Tables 1, 2 and 3 are directly interpretable for patients, although many of them are. A patient with diabetes will know what are preferred insulin and glucose values. The descriptions furthermore provide interpretable explanations to doctors, and they may furthermore be well aware to what part of the distribution these particular subgroups belong (i.e., whether certain values should be considered high or low).

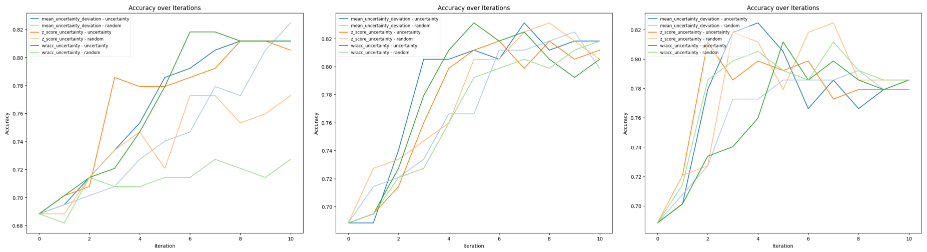
6.2 Evaluating EMM as an AL method

Next, we investigate different strategies for sample selection (random vs based on uncertainty score), as well as how the number of samples within the group and the number of iterations impact the results.

Figure 1 displays the increase in prediction accuracy from $t = 0$ to $t = 10$ (x-axis) when adding (a) 1 sample, (b) 2 samples, and (c) 5 samples per subgroup per iteration (for a total of $q = 10$ subgroups). We compare the result for three quality measures (φ_{mud} in blue, φ_{mudef} in green, φ_{uz} in orange) and compare adding samples at random (low density color) with adding samples based on uncertainty score (highest uncertainty first, in bold).

The figure shows that the accuracy of the classifier increases significantly in the first couple of iterations, even though only few samples have been added; 40 and 80 samples in the first four iterations when adding 1 or 2 samples per iteration respectively. This constitutes less than 10% and less than 20% of the full dataset. Furthermore, Figure 1 demonstrates that the uncertainty sampling technique performs better than the random sampling technique. This is to be expected. However, there is no clear difference between the performance of the three quality measures. Interestingly, Figure 1c shows that when we have labeled sufficient number of samples, the differences between the two sampling

Fig. 1: Classifier accuracy (y-axis) when using EMM as an AL technique. We compare 3 quality measures (φ_{mud} in blue, φ_{mudef} in green, φ_{uz} in orange), 2 sample selection methods (random with low density, uncertain samples first, in bold) and evaluate differences in adding 1, 2 or 5 samples per subgroup per iteration, for 10 iterations (x-axis) and $q = 10$ subgroups).



(a) 1 sample (b) 2 samples (c) 5 samples

techniques disappear (i.e., the light and bold lines overlap). Here we can also observe that when a big proportion of the dataset has been included in the training dataset, the accuracy decreases, or at least no longer increases. This occurs in Figure 1c around $t = 4$, when about $5 \cdot 10 \cdot 4 = 200$ samples have been added already. This finding is a sign that the model is overfitting, or at least does not gain any additional knowledge from the added samples.

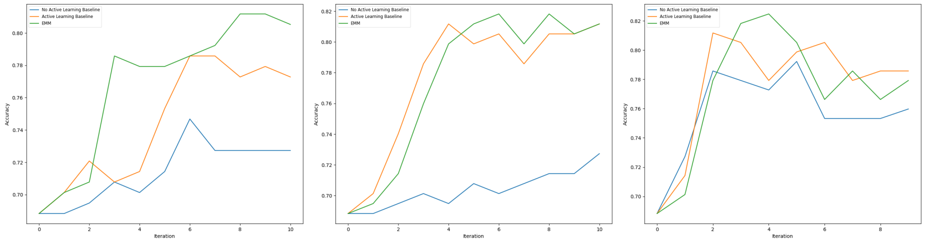
6.3 Comparison against AL baseline

We now compare our EMM-based AL method against two baselines: the null baseline where samples are randomly selected from U without further oracle information (in blue in Figure 2) and an AL baseline that performs uncertainty sampling (in orange). We perform our EMM-based AL method using quality measure φ_{mud} ((in green), since all quality measures performed similarly well on the previous analyses in Section 6.2, we proceed to use the Mean Uncertainty Deviation as quality measure due to its good qualitative performance on the dataset in Section 6.1. That is, φ_{mud} resulted in fairly smaller datasets, which increases our hopes that the subgroups are more diverse and cover various parts of the feature space.

Clearly, Figure 2 demonstrates that an AL paradigm improves the performance of the classification model. This is particularly apparent when adding 2 samples per iteration (center).

Furthermore, we observe that when adding 1 sample per subgroup per iteration, our method performs better than the AL baseline. This can be seen from the green line, that results in severely higher accuracy values from $t = 3$ onward. These findings support our hypothesis that using EMM is beneficial for selecting the set of most informative samples. Compared to selecting samples based on just the uncertainty values, as does the AL baseline, our EMM-based search is

Fig. 2: Classifier accuracy (y-axis) when using EMM as an AL technique. Comparison against a null baseline (adding random samples from U without further oracle information, in blue) and an AL baseline that performs uncertainty sampling, in orange. We perform our EMM-based AL method (in green) using quality measure φ_{mud}, and evaluate differences between adding 1, 2 or 5 samples per subgroup per iteration, for 10 iterations (x-axis) and $q = 10$ subgroups).



(a) 1 sample

(b) 2 samples

(c) 5 samples

able to discover a larger variety of subgroups covering more parts of the feature space. Especially when only few samples are added to the training data, our EMM-based approach outperforms the AL baseline method.

When adding two samples per iteration, our method performs similarly to the AL baseline; when adding sufficiently large number of samples to the training set, the differences between an EMM-based AL method, an AL baseline and a random, null baseline disappear. Here, there are so many samples selected that almost all samples in U will end up in D_t. Similar as in Figure 1, the accuracy starts decreases after several iterations; possibly due to overfitting.

7 Discussion and Conclusion

In this paper, we proposed a novel method for interpretable sample selection in the Active Learning (AL) process using Exceptional Model Mining (EMM). Traditional AL methods are prone to oversampling certain regions in the feature space. As an alternative, we use the EMM framework to discover a diverse list of subgroups of informative samples. By integrating EMM, our approach identifies subgroups within the unlabeled data whose uncertainty characteristics deviate from the rest of the data. By choosing samples from within each identified group, the training set includes more diverse and representative samples.

Our experiment on the Pima Indians Diabetes dataset show that EMM is capable of identifying descriptive groups within the unlabeled set which allows us to make a more informed selection of the samples for labeling. By labeling the most uncertain samples from the discovered subgroups, we show that the EMM-based method performs better than selecting random samples for labeling, especially when the number of selected samples is small (e.g., 1 sample per subgroup; 10 samples per iteration). Currently, our method is on par with uncertainty based AL when more samples are added. We hypothesize that our method could further outperform existing AL methods also in these scenarios, when our beam search is enhanced with more anti-redundancy techniques, following [38]. The strength of our proposed method is to create a diverse list of subgroups such that a diverse set of samples can be added.

At first instance, it seems that the choice of quality measure has little impact on the performance of our method. However, similar as in traditional AL where the sampling mechanism is increasingly improved, we hypothesize that a more informed quality measure could further improve our proposed method.

One limitation of AL methods which we do not address is the computational cost. In our proposed method this is even higher since both the classification model and the EMM model need to be run at each iteration. Future work could focus on developing more efficient algorithms or heuristics for subgroup discovery and EMM evaluation. It would also be interesting to see whether our proposed EMM-based AL method could outperform existing AL baseline without re-running the search algorithm during every iteration. For instance, we could run the subgroup search once at $t = 0$, and then iteratively pick samples from these subgroups. This approach would significantly reduce the computational cost of our proposed EMM-based AL method.

In sum, this paper presents an EMM-based AL approach that provides interpretable explanations regarding selected, informative samples, and additionally is a standalone, AL method that gives good classification performance, especially when labeling budget is low (and few samples are selected). At a larger scale, we demonstrate that EMM is a valuable framework that could serve as an explainability method in larger AI systems. Moreover, we demonstrate that the list of discovered subgroups can be utilized for more than knowledge discovery.

Acknowledgments. We gratefully acknowledge the continuous support of Dr. Wouter Duivesteijn and Prof. Mykola Pechenizkiy, in particular the way they designed the Data Science and AI course on Research Topics in Data Mining and provide an opportunity for Master students to contribute to methodological advancements. Another part of this work is executed in the EDIC project as part of the Data2People program financed by the Dutch Research Council (NWO).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahmad, M.A., Eckert, C., Teredesai, A.: Interpretable machine learning in healthcare. In: Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics. pp. 559–560 (2018)
2. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion **58**, 82–115 (2020)

3. Barz, M., Sonntag, D.: Incremental improvement of a question answering system by re-ranking answer candidates using machine learning. In: Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems. pp. 367–379. Springer (2021)
4. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine Learning **15**(2), 201–221 (May 1994). https://doi.org/10.1007/BF00993277, `https://doi.org/10.1007/BF00993277`
5. Culotta, A., McCallum, A.: Reducing labeling effort for structured prediction tasks. In: AAAI. vol. 5, pp. 746–751 (2005)
6. Dilsizian, S.E., Siegel, E.L.: Artificial intelligence in medicine and cardiac imaging: Harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. Current Cardiology Reports **16**(1), 441 (Jan 2013). https://doi.org/10.1007/s11886-013-0441-8, `https://doi.org/10.1007/s11886-013-0441-8`
7. Donoso-Guzmán, I., Ooge, J., Parra, D., Verbert, K.: Towards a comprehensive human-centred evaluation framework for explainable AI. In: World Conference on Explainable Artificial Intelligence. pp. 183–204. Springer (2023)
8. Du, X., Duivesteijn, W., Pechenizkiy, M.: Conformalized exceptional model mining: Telling where your model performs (not) well. In: Proc. ECML PKDD (2025, to appear)
9. Duivesteijn, W., Feelders, A.J., Knobbe, A.: Exceptional model mining. Data Mining and Knowledge Discovery **30**(1), 47–98 (Jan 2016). https://doi.org/10.1007/s10618-015-0403-4, `https://doi.org/10.1007/s10618-015-0403-4`
10. Duivesteijn, W., Thaele, J.: Understanding where your classifier does (not) work–the SCaPE model class for EMM. In: Proc. ICDM. pp. 809–814 (2014)
11. Göpfert, J.P., Kuhl, U., Hindemith, L., Wersing, H., Hammer, B.: Intuitiveness in active teaching. IEEE Transactions on Human-Machine Systems **52**(3), 458–467 (2021)
12. Herrera, F., Carmona, C.J., González, P., Del Jesus, M.J.: An overview on subgroup discovery: Foundations and applications. Knowledge and Information Systems **29**(3), 495–525 (2011)
13. Huang, S.J., Jin, R., Zhou, Z.H.: Active learning by querying informative and representative examples. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(10), 1936–1949 (2014). https://doi.org/10.1109/TPAMI.2014.2307881
14. Hudson, D., Wiltshire, T.J., Atzmueller, M.: Local exceptionality detection in time series using subgroup discovery: An approach exemplified on team interaction data. In: Proc. DS. pp. 435–445 (2021)
15. Iferroudjene, M., Lonjarret, C., Robardet, C., Plantevit, M., Atzmueller, M.: Methods for explaining top-N recommendations through subgroup discovery. Data Mining and Knowledge Discovery **37**(2), 833–872 (2023)
16. Klosgen, W.: Subgroup mining. In: Computational Intelligence in Data Mining. pp. 39–49. Springer Vienna, Vienna (2000)
17. Leitner, J., Chiang, P.H., Agnihotri, P., Dey, S.: The effect of an AI-based, autonomous, digital health intervention using precise lifestyle guidance on blood pressure in adults with hypertension: Single-arm nonrandomized trial. JMIR cardio **8**, e51916 (2024)
18. Lemmerich, F., Atzmueller, M., Puppe, F.: Fast exhaustive subgroup discovery with numerical target concepts. Data Mining and Knowledge Discovery **30**(3), 711–762 (2016)

19. Lewis, D., Gale, W.: A sequential algorithmfor training text classifiers. In: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University. pp. 3–12 (1994)
20. Liu, W., Dai, B., Humayun, A., Tay, C., Yu, C., Smith, L.B., Rehg, J.M., Song, L.: Iterative machine teaching. In: International Conference on Machine Learning. pp. 2149–2158. PMLR (2017)
21. Marcu, A.M., Chen, L., Hünermann, J., Karnsund, A., Hanotte, B., Chidananda, P., Nair, S., Badrinarayanan, V., Kendall, A., Shotton, J., et al.: LingoQA: Visual question answering for autonomous driving. In: European Conference on Computer Vision. pp. 252–269. Springer (2024)
22. McCallum, A.K., Nigam, K., et al.: Employing em and pool-based active learning for text classification. In: ICML. vol. 98, pp. 350–358. Citeseer (1998)
23. Melville, P., Yang, S.M., Saar-Tsechansky, M., Mooney, R.: Active learning for probability estimation using jensen-shannon divergence. In: 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005. Proceedings 16. pp. 268–279. Springer (2005)
24. Mondal, I., Ganguly, D.: Alex: Active learning based enhancement of a model's explainability (2020), `https://arxiv.org/abs/2009.00859`
25. Murray, C., Allingham, J.U., Antorán, J., Hernández-Lobato, J.M.: Addressing bias in active learning with depth uncertainty networks... or not (2021). https://doi.org/2112.06926
26. Nass, C., Moon, Y.: Machines and mindlessness: Social responses to computers. Journal of social issues **56**(1), 81–103 (2000)
27. Nath, V., Yang, D., Roth, H.R., Xu, D.: Warm start active learning with proxy labels and selection via semi-supervised fine-tuning. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 297–308. Springer Nature Switzerland, Cham (2022)
28. Nguyen, H.T., Smeulders, A.: Active learning using pre-clustering. In: Proceedings of the twenty-first international conference on Machine learning. p. 79 (2004)
29. Pieters, B.F., Knobbe, A., Dzeroski, S.: Subgroup discovery in ranked data, with an application to gene set enrichment. In: Proc. Preference Learning Workshop at ECML PKDD. pp. 1–18 (2010)
30. Scheffer, T., Decomain, C., Wrobel, S.: Active hidden markov models for information extraction. In: International symposium on intelligent data analysis. pp. 309–318. Springer (2001)
31. Schouten, R.M.: Exceptional Model Mining for Hierarchical Data. PhD thesis (2025)
32. Schouten, R.M., Duivesteijn, W., Pechenizkiy, M.: Exceptional Model Mining for Repeated Cross-Sectional Data (EMM-RCS). In: Proc. SDM. pp. 585–593 (2022)
33. Settles, B.: Active learning literature survey. Tech. rep. (2009)
34. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the 2008 conference on empirical methods in natural language processing. pp. 1070–1079 (2008)
35. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proceedings of the fifth annual workshop on Computational learning theory. pp. 287–294 (1992)
36. Tharwat, A., Schenck, W.: A survey on Active Learning: State-of-the-art, practical challenges and research directions. Mathematics **11**(4), 820 (2023)
37. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. Journal of machine learning research **2**(Nov), 45–66 (2001)

38. Van Leeuwen, M., Knobbe, A.: Diverse subgroup set discovery. Data Mining and Knowledge Discovery **25**, 208–242 (2012)
39. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: Proc. PKDD. pp. 78–87 (1997)
40. Zhu, X.: Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In: Proceedings of the AAAI conference on artificial intelligence. vol. 29 (2015)