Minority Groups Detection

1 Motivation

- The use of "found data" is very common. A major problem in the use of "found data" is that it may not be adequate to the user's intended use. In particular it may have underrepresented groups
- Identifying underrepresented groups was done in [1], but the assumption that those minority groups always have lower accuracy is incorrect.
- Identifying underrepresented groups with low accuracy may require prior domain knowledge and exhaustive search.
- Given such a minority group (with inadequate accuracy compared to the overall model accuracy), the ideal way to eliminate the problem, is by adding more data on this group. However, this task may be expensive, and even impossible, particularly in the case of "found data".
- We propose alternative method of data adjustment to tackle the problem:
  - data re-weighting
  - synthetically generated tuples based on the tuples in the data
- We identify and characterize the cases where our alternative solutions are applicable and demonstrate their usefulness in those cases

2 Problem Definition

Intuitively, given a ML model, and training/testing data, our objective is

- Automatically identify underrepresented groups with low performance in the data
- Determine the feasibility of data adjustment for a given minority group

We assume the data is represented using a single relational database, and that the relation's attributes values are categorical. [[[**(Yuval) maybe the categorical assumption can be relaxed**]]]

Definition 2.1 (Patterns) *Let D be a database with attributes $\mathcal{A} = \{A_1, \ldots, A_n\}$ and let $Dom(A_i)$ be the active domain of A_i for $i \in [1..n]$. A* pattern *p over D is set of $\{A_{i_1} = a_1, \ldots, A_{i_k} = a_k\}$ where $\{A_{i_1}, \ldots, A_{i_k}\} \subseteq \mathcal{A}$ and $a_j \in Dom(A_{i_j})$ for each A_{i_j} in p. We use $Attr(p)$ to denote the set of attributes in p.*

We say that a tuple $t \in D$ satisfies *a pattern p and denote it by $t \vDash p$ if $t.A_i = a_i$ for each A_i appearing in p.*

Definition 2.2 *Let D be a database and M be a classifier. Denoting the model prediction of M on $t \in D$ with $M(t)$ and the t's true class with $C(t)$, we define $f_M : D \mapsto \{0, 1\}$ as follows.*

$$f_M(t) = \begin{cases} 1 & \textit{if } C(t) = M(t) \\ 0 & \textit{otherwise} \end{cases}$$

Definition 2.3 (Pattern Accuracy) *Given a model M a database D and a pattern p, the* accuracy *of p is*

$$acc_M(p) = \frac{\sum_{t \in D. t \vDash p} f_M(t)}{|\{t \in D \mid t \vDash p\}|}$$

Definition 2.4 *Given a model M a database D and an error threshold τ, we say that a p is* most general *pattern with accuracy below τ if $acc_M(p) < \tau$ and $\forall p' \subsetneq p \; acc_M(p') \geq \tau$*

Problem 2.5 (Low Accuracy Patterns Detection) *Given a database D, a model M an accuracy threshold τ_a and a size threshold τ_s, find all most general patterns with count $\geq \tau_s$ and accuracy below τ_a*

[[[(Yuval) TBD: hardness?]]]

Algorithm

Algorithm 1: Low Accuracy Pattern Detection

```
input  : A database $D$ a model $M$ an accuracy threshold $\tau_a$ and a size
         threshold $\tau_s$.
output: $\mathcal{P} = \{p_1, \ldots, p_n\}$ s.t. $\forall p_i \in \mathcal{P}$ $p_i$ is a most general pattern with
         $c_D(p_i) \geq \tau_s$ and $acc_M(p_i) < \tau_a$.
$\mathcal{P} \leftarrow 0$
$miss \leftarrow \{t \in D | f_M(t) = 0\}$        miss-classified
$cands \leftarrow$ generalPatterns $(miss, \tau_s \cdot \tau_a)$
for $p \in cands$ do
    if $acc_M(p) < \tau_a$ then        and $C_D(p) \geq \tau_s$
        $\mathcal{P} \leftarrow \mathcal{P} \cup \{p\}$
return $\mathcal{P}$
```

$50 \times 0.6 = 30$

$40 \times 0.7 = 28$

- $miss_M(D)$ is the dataset obtained from D by removing all tuples that were classified correctly.
- `generalPatterns`(D, S) finds most general patterns in D with count $> S$. This can be done by traversing a lattice, similarly to the algorithms in [1]
- the computation of $acc_M(p)$ in line 5 can be optimized. Given $c_D(p)$ (this should be part of the computation of the `generalPatterns` procedure), it is enough to compute $c_{miss}(p)$ to determine the value $acc_M(p)$.

Other problems

Problem 2.6 (Data Adjustment Feasibility) *Given a database D, a model M and a most general pattern with low accuracy, determine whether data adjustment method can be used to improve the pattern accuracy.*

[[[**(Yuval) needs to bee more formally defined. E.g., given two threshold τ_1 and τ_2 we want $acc_M(p) > tau_1$ but the overall accuracy to be above τ_2**]]]

Possible criteria

- Learning based on the minority results in higher accuracy
- Diversity measures

[[[**(Yuval) Another variant of the Low Accuracy Pattern Detection: instead of a classifier we are given a scoring function and we want to find groups that have a large difference between the predicted score and the real score**]]]

References

[1] A. Asudeh, Z. Jin, and H. V. Jagadish. Assessing and remedying coverage for a given dataset. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*, pages 554–565. IEEE, 2019.