Technical Report on "Tailoring the Shapley Value for In-context Example Selection towards Data Wrangling"

Zheng Liang[1], Hongzhi Wang[1,✉], Xiaoou Ding[1], Zhiyu Liang[1], Chen Liang[1], Jianzhong Qi[2]
[1] School of Computer Science, Harbin Institute of Technology, Harbin, China
[2] Faculty of Engineering and Information Technology, University of Melbourne, Melbourne, Australia
{lz20, wangzh, dingxiaoou,zyliang,chenliang}@hit.edu.cn, jianzhong.qi@unimelb.edu.au

I. Proof for the Propositions

Proposition 1. *(Constrained Shapley Value Uniqueness): For any "game" (D, U), where U is a utility function that maps a subset S of players $D = \{d_1, d_2, \ldots, d_n\}$ to a real number: $U(S) \rightarrow \mathbb{R}$, if U can only take coalitions (i.e., subset S of D) containing at most k players (i.e., candidate examples) as input, $CSV(\cdot)$ is the* only *value function $F(d_i)$ that satisfies the following properties for reward allocation:*

Symmetry: Any two candidate examples with equal marginal contributions to every subset S receive the same reward. Formally, $\forall d_i, d_j \in D$, if $\forall S \subset D, |S| < k : U(S \cup \{d_i\}) = U(S \cup \{d_j\})$, then $F(d_i) = F(d_j)$, where $F(d_i)$ and $F(d_j)$ are the rewards of d_i and d_j.

Additivity: For any subset $S \subset D, |S| \leq k$, the utility function value $U(S)$ can be fully divided among the candidate examples, i.e., $\forall S \subset D, |S| \leq k : U(S) = \sum_{d_i \in S} F(d_i)$.

Balance: For any player $d_i \in D$ playing any two games (D, F_1) and (D, F_2) getting reward $F_1(d_i)$ and $F_2(d_i)$, respectively; its reward allocation for the game $(D, F_1 + F_2)$ is $F_1(d_i) + F_2(d_i)$.

Zero element: A candidate example with zero contribution to the reward of every subset of D with up to k elements has a reward of 0. Formally, $\forall d_i \in D$, if $\forall S \subset D, |S| < k : U(S \cup \{d_i\}) = U(S)$, then $F(d_i) = 0$.

Proof. We show the uniqueness of the CSV based on its definition. Note that this is our corollary of the theorem in a previous work [1]. We give this proof to make our paper self-contained.

We use $r, s, n, \ldots$ to represent the size of sets $R, S, N, \ldots$, respectively. The sets will be introduced below when they are needed.

Let P be the universe of players. Define a game to be any set function $v : P \rightarrow \mathbb{R}$ that maps from a subset of U to a real number, where a superadditive game satisfies:

1) $v(\emptyset) = 0$;
2) $v(S) \geq v(S \cap T) + v(S - T)$, $\forall S, T \subset U \wedge |S| \leq k \wedge |T| \leq k$;
3) A carrier of v is any subset $N \subset U$ with $v(S) = v(N \cap S), \forall S \subset U$. $F_i[v] = 0$ for **zero elements** $\forall i \in S \setminus N$;

✉ Hongzhi Wang is the corresponding author.

Our goal is to compute $F_i[v]$, the valuation of i in game v, which is supposed to satisfy the four properties. Following the previous work [**?**], we compute $F_i[v]$ in three steps:

Step 1: Decompose v into the weighted sum of certain symmetric games.

Step 2: Compute the weight and the valuation function in the symmetric games.

Step 3: Compute $F_i[v]$.

Step 1: We first consider certain symmetric games. For any $R \subset U, R \neq \emptyset$, we define v_R:

$$v_R(S) = \begin{cases} 0 & \text{if } R \subset S, |S| \leq k \\ 1 & \text{if } R \not\subset S, |S| \leq k \\ 0 & \text{if } |S| > k \end{cases} \tag{1}$$

A immediate corollary to the **Additivity property** is that $F[v - w] = F[v] - F[w]$ if v, w, and $v - w$ are all games. Therefore, according to the previous work [**?**], any game v is a linear combination of symmetric games v_R:

$$v = \sum_{R \subset N, R \neq \emptyset} c_R(v) v_R, \tag{2}$$

where the coefficients are given by

$$c_R(v) = \sum_{T \subset R} (-1)^{r-t} v(T) \tag{3}$$

Step 2: Suppose a projection $\pi : R \rightarrow R, \pi(i) = j$, also $\pi(R) = R$, by the **Balance property**, we have:

$$F_i[v_R] = F_{\pi(i)}[v_{\pi(R)}] = F_j[v_R]. \tag{4}$$

Further, based on the **Symmetry property**, we have:

$$1 = v_R(R) = \sum_{j \in R} F_j[v_R] = r F_i[v_R]. \tag{5}$$

Therefore,

$$F_i[v_R] = \begin{cases} \frac{1}{r} & \text{if } i \in R, \\ 0 & \text{if } i \notin R. \end{cases} \tag{6}$$

Step 3: We now apply (19) to (15) and obtain:

$$F_i[v] = \sum_{R\subset N, i\in R} \frac{c_R(v)}{r}, \forall i \in N \quad (7)$$

$$= \sum_{R\subset N, i\in R} \frac{\sum_{T\in R}(-1)^{r-t}v(T)}{r}, \forall i \in N \quad (8)$$

$$= \sum_{S\subset N, i\in S, |S|\le k} \frac{(s-1)!(n-s)!}{n!} v(S) \quad (9)$$

$$- \sum_{S\subset N, i\notin S, |S|\le k-1} \frac{(s)!(n-s-1)!}{n!} v(S), \forall i \in N \quad (10)$$

$$= CSV(d_i) \quad (11)$$

Therefore, a unique value function F satisfying Balance, Symmetry, and Additivity, for games with finite carriers, is given by the definition of the Constrained Shapley Value. □

Proposition 2. *(Monte Carlo Marginal Contribution Approximation Quality [**?**]) According to Hoeffding's inequality, given the range $r = max(CSV(d_i)) - min(CSV(d_i))$ of $CSV(d_i)$, a CSV estimation error bound ϵ, and a confidence level $1-\delta$, Algorithm **??** takes $\frac{mr^2 avl(D)k^2}{4\epsilon^2}\log\frac{2n}{\delta}$ API token costs and $\frac{mr^2}{2\epsilon^2}k\log\frac{2n}{\delta}$ queries to an LLM to ensure $P(|\overline{CSV(d_i)} - CSV(d_i)| \ge \epsilon) \le \delta$, where $avl(D)$ denotes the average number of tokens in serialized EM examples.*

Proof. For any random variable $S_{min} \le S \le S_{max}$, according to the Hoeffding's inequality we have:

$$P(|S-E(S)| \ge t) \le 2\cdot\exp\left(-\frac{2t^2}{\sum_{i=1}^{m}(S_{max}-S_{min})^2}\right) \quad (12)$$

For $\forall d_i \in D$, let $S = \sum_{i=1}^{n} CSV(d_i)$, r be the difference of the maximum and minimum values of $CSV(d_i)$, c be the number of 'permutations' [**?**] in Line 3 of Algorithm **??**, the Hoeffding's inequality entails that:

$$P(|S - cE(S)| \ge t) = P(|CSV(d_i) - \overline{CSV(d_i)}| \ge \frac{t}{c}) \quad (13)$$

$$P(|CSV(d_i) - \overline{CSV(d_i)}| \ge \epsilon) \le 2\cdot\exp\left(-\frac{2c^2\epsilon^2}{cr^2}\right) \quad (14)$$

Our aim is to make the right hand side to be at most δ:

$$2\cdot\exp\left(\frac{-2c\epsilon^2}{r^2}\right) \le \delta \quad (15)$$

$$c \ge \frac{r^2\cdot\log\frac{2}{\delta}}{2\epsilon^2} \quad (16)$$

As each sample can only be used by one CSV approximation, the total number of utility function computation is $\frac{nc}{k} \ge \frac{nr^2\cdot\log\frac{2}{\delta}}{2k\epsilon^2}$. Each permutation requires examining k candidate examples, which means km questions to the LLM. Thus, the number of QA turns is at least $\frac{mnr^2\cdot\log\frac{2}{\delta}}{2\epsilon^2}$. On average, $\frac{k}{2}$ candidate examples are used for LLM in-context prompting, and hence Algorithm **??** consumes $\frac{avl(D)kmnr^2\cdot\log\frac{2}{\delta}}{4\epsilon^2}$ API tokens. This completes the proof. □

Proposition 3. *(Effectiveness of Activated Contribution Approximation) Given a set of candidate examples $D = \{d_1, d_2, \dots, d_n\}$, the constrained Shapley value of d_i can be computed by:*

$$CSV(d_i) = \frac{1}{n}\sum_{S\subset D, 0\le|S|\le k} \frac{AC(S,d_i)}{\binom{n-1}{|S|}} \quad (17)$$

Proof. We rewrite the definition of CSV in Equation **??** to:

$$\begin{aligned} CSV(d_i) &= \frac{1}{n}\sum_{\substack{S\subset D\backslash\{d_i\},\\ 0\le|S|\le k-1}} \frac{U(S\cup\{d_i\}) - U(S)}{\binom{n-1}{|S|}} \\ &= \frac{1}{n}\sum_{\substack{S\subset D\backslash\{d_i\},\\ 0\le|S|\le k-1}} \frac{U(S\cup\{d_i\})}{\binom{n-1}{|S|}} - \frac{1}{n}\sum_{\substack{S\subset D\backslash\{d_i\},\\ 0\le|S|\le k-1}} \frac{U(S)}{\binom{n-1}{|S|}} \end{aligned} \quad (18)$$

Let $S' = S\cup\{d_i\}$. Since $d_i \notin S$, we have $S = S'\backslash\{d_i\}$. Putting S' into the equation above yields:

$$CSV(d_i) = \frac{1}{n}\sum_{\substack{d_i\in S',\\ 1\le|S'|\le k,\\ S'\subset D}} \frac{U(S')}{\binom{n-1}{|S'|-1}} - \frac{1}{n}\sum_{\substack{d_i\notin S,\\ 0\le|S|\le k-1,\\ S\subset D}} \frac{U(S)}{\binom{n-1}{|S|}} \quad (19)$$

Next, we use a shared variable S^* to replace S' and S. Since $S' \neq S$ always holds, they can be viewed as two cases of variable S^*.

$$\begin{aligned} CSV(d_i) &= \frac{1}{n}\sum_{\substack{d_i\in S^*,\\ 1\le|S^*|\le k,\\ S^*\subset D}} \frac{U(S^*)}{\binom{n-1}{|S^*|-1}} - \frac{1}{n}\sum_{\substack{d_i\notin S^*,\\ 0\le|S^*|\le k-1,\\ S^*\subset D}} \frac{U(S^*)}{\binom{n-1}{|S^*|}} \\ &= \frac{1}{n}\sum_{\substack{d_i\in S^*,\\ 1\le|S^*|\le k,\\ S^*\subset D}} \frac{U(S^*)\cdot(\frac{n}{|S^*|}-1)}{\binom{n-1}{|S^*|}} - \frac{1}{n}\sum_{\substack{d_i\notin S^*,\\ 0\le|S^*|\le k-1,\\ S^*\subset D}} \frac{U(S^*)}{\binom{n-1}{|S^*|}} \end{aligned} \quad (20)$$

According to the definition of Activated Contribution and the fact that $|S^*| \neq 0$ when $d_i \in S^*$, we can rewrite the weight term into a unified weight as follows.

$$\begin{aligned} CSV(d_i) &= \frac{1}{n}\sum_{\substack{d_i\in S^*,\\ 0\le|S^*|\le k,\\ S^*\subset D}} \frac{U(S^*)f(S^*,d_i)}{\binom{n-1}{|S^*|}} + \frac{1}{n}\sum_{\substack{d_i\notin S^*,\\ 0\le|S^*|\le k-1,\\ S^*\subset D}} \frac{U(S^*)f(S^*,d_i)}{\binom{n-1}{|S^*|}} \\ &= \frac{1}{n}\sum_{\substack{0\le|S^*|\le k,\\ S^*\subset D}} \frac{U(S^*)f(S^*,d_i)}{\binom{n-1}{|S^*|}} - \frac{1}{n}\sum_{\substack{d_i\notin S^*,\\ |S^*|=k,\\ S^*\subset D}} \frac{U(S^*)f(S^*,d_i)}{\binom{n-1}{|S^*|}} \\ &= \frac{1}{n}\sum_{\substack{0\le|S^*|\le k,\\ S^*\subset D}} \frac{AC(S^*,d_i)}{\binom{n-1}{|S^*|}} \end{aligned} \quad (21)$$

Replacing S^* with S completes the proof.

□

Proposition 4. *(Unbiased Estimation of Activated Contribution Approximation) Given a set of players $D = \{d_1, d_2, \dots, d_n\}$, Algorithm **??** gives an unbiased estimation of the constrained Shapley value for every player, that is, $E(\overline{CSV(d_i)}) = CSV(d_i), 1 \le i \le n$.*

Proof. Let $CSV_{i,j}$ be the CSV of a coalition of size j calculated as follows:

$$CSV_{i,j} = \frac{1}{n} \sum_{\substack{S \subset D,\\ |S|=j}} \frac{AC(S, d_i)}{\binom{n-1}{|S|}} \tag{22}$$

By the definition of CSV, we have the following immediately:

$$\begin{aligned} CSV(d_i) &= \frac{1}{n} \sum_{1 \le j \le n} CSV_{i,j} \\ CSV_{i,j} &= E(AC(S, d_i)) \end{aligned} \tag{23}$$

In Algorithm **??**, all possible S is sampled from N (Line 3) with sample allocation. Thus, according to Theorem 4.5 of a previous work [**?**], $\frac{\overline{CSV_{i,j}}}{m_{i,j}}$ is an unbiased estimation of $AC(S, d_i)$. Therefore, the proposition holds.

$$CSV(d_i) = \sum_{j=1}^{n} CSV_{i,j} = \sum_{j=1}^{n} E(AC(S, d_i)) = \sum_{j=1}^{n} E(\frac{\overline{CSV_{i,j}}}{m_{i,j}}) = E(\overline{CSV_i}) \tag{24}$$

□

Proposition 5. *(AC-based Minimized Deviation Approximation Quality) With sample allocation towards deviation minimization, Algorithm **??** takes $\frac{2r^2 \log \frac{2}{\delta} avl(D)\sqrt{n}}{\epsilon^2} \sum_{j=1}^{k} \frac{(j+1)}{\sqrt[3]{j}}$ API token costs to ensure $P(|\overline{CSV(d_i)} - CSV(d_i)| \ge \epsilon) \le \delta$.*

Proof. According to the optimal solution to the relaxed Deviation Minimization problem, within the probability of at least $1 - \delta$, we have:

$$|\overline{CSV(d_i)} - CSV(d_i)| \le 2r \sqrt{\frac{\log \frac{2}{\delta} \cdot avl(D) \cdot n}{2B} \sum_{j=1}^{k} \frac{(j+1)}{\sqrt[3]{j}}} \tag{25}$$

By setting the right hand side as ϵ, we have:

$$B = \frac{2r^2 \log \frac{2}{\delta} avl(D)\sqrt{n}}{\epsilon^2} \sum_{j=1}^{k} \frac{(j+1)}{\sqrt[3]{j}} = O(\frac{mk^2\sqrt{n}}{\epsilon^2} \log \frac{1}{\delta}) \tag{26}$$

□

Proposition 6. *(AC-based Regret Minimizing Approximation Quality) With sample allocation towards regret minimization, the error probability of Algorithm **??** satisfies the following inequality:*

$$e_n = P(\cup_{i \le k \le j} CSV(d_i) < CSV(d_j)) \le 2k^2 exp(-\frac{n-k}{8\overline{\log}k \cdot H}),$$

where $H = max_{i \in \{1,\dots,K\}} i \cdot (|CSV(d_i) - CSV(d_{i+1})|)^{-2}$ and $\overline{\log}K = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$.

Proof. Consider the event ξ defined by

$$\xi = \{j \in \{1, ..., K\}, |\frac{1}{n_k} \sum_{s=1}^{n_k} X_s - f_j| \le \frac{1}{2}\Delta_{K+1-k}\}. \tag{27}$$

By Hoeffding's Inequality and with the sample allocation strategy in Equation 13, the probability of the complementary event $\bar{\xi}$ can be bounded as follows:

$$\begin{aligned} P(\bar{\xi}) &\le \sum_{j=1}^{K} \sum_{k=1}^{K-1} P(|\frac{1}{n_k} \sum_{s=1}^{n_k} X_s - f_j| \le \frac{1}{2}\Delta_{K+1-k}) \\ &\le \sum_{j=1}^{K} \sum_{k=1}^{K-1} 2exp(2n_k(\Delta_{K+1-k})/2)^2) \\ &\le 2K^2 exp(-\frac{n-K}{2\overline{\log}K \cdot H}). \end{aligned} \tag{28}$$

Here, the last inequality comes from the fact that:

$$\begin{aligned} & n_k(\Delta_{K+1-k})^2 \\ &\ge \frac{n-K}{\overline{\log}(K)(K+1-H)(\Delta_{K+1-k})^{-2}} \\ &\ge \frac{n-K}{\overline{\log}(K) \cdot H}. \end{aligned} \tag{29}$$

Thus, it suffices to show that on event ξ, the algorithm makes no error. We prove this by induction on k. Let $k \ge 1$. Assume that the algorithm makes no error in all previous $k-1$ stages, i.e., no bad arm $\mu_i < \theta$ has been accepted and no good arm $\mu_i \ge \theta$ has been rejected. Event ξ implies that at the end of stage k, all empirical means are within $\frac{1}{2}(\Delta_{K+1-k})^{-2}$ of the respective true means.

Let $A_k = \{a_1, \dots, a_{K+1-k}\}$ be the set of active arms during stage k. We order the a_i's such that $\mu_{a_1} > \mu_{a_2} > \dots > \mu_{a_{K+1-k}}$. Let $m' = m(k)$ be the number of arms left to find in stage k. The fact that no error has occurred in the first $k-1$ stages implies:

$$a_1, a_2, \dots, a_{m'} \in \{1, \dots, m\} \tag{30}$$

and

$$a_{m'+1}, \dots, a_{K+1-k} \in \{m+1, \dots, K\} \tag{31}$$

If an error is made at stage k, it can be one of the following two types:

(1) The algorithm accepts a_j at stage k for some $k \ge m'+1$.

(2) The algorithm rejects a_j at stage k for some $j \le m'$.

Let $\sigma = \sigma_k$ be the bijection (from $\{1, \dots, K+1-k\}$ to A_k) such that $\overline{\mu}_{\sigma(1),n_k} \ge \overline{\mu}_{\sigma(2),n_k} \ge \dots \ge \overline{\mu}_{\sigma(K+1-k),n_k}$. Suppose Type 1 error has occurred. Then $a_j = \sigma(1)$, since if the algorithm accepts, it must accept the empirical best arm. Furthermore, we have:

$$\overline{\mu}_{a_j,n_k} - \theta \ge \theta - \overline{\mu}_{\sigma(K+1-k),n_k}, \tag{32}$$

since otherwise the algorithm would rather reject arm $\sigma(K+1-k)$. The condition $a_j = \sigma(1)$ and the event ξ implies that:

$$\begin{aligned} \overline{\mu}_{a_j,n_k} &\geq \overline{\mu}_{a_j,n_k}, \\ \mu_{a_j} + \frac{1}{2}(\Delta_{K+1-k}) &\geq \mu_{a_1} - \frac{1}{2}(\Delta_{K+1-k}), \\ (\Delta_{K+1-k}) &\geq \mu_{a_1} - \mu_{a_j} \geq \mu_{a_1} - \theta \end{aligned} \tag{33}$$

We then look at Condition (40). In the event of ξ, for all $i \leq m'$, we have:

$$\begin{aligned} \overline{\mu}_{a_j,n_k} &\geq \mu_{a_j} - \frac{1}{2}\Delta_{(K+1-k)} \\ &\geq \mu_{a_{m'}} - \frac{1}{2}\Delta_{(K+1-k)} \\ &\geq \theta - \frac{1}{2}\Delta_{(K+1-k)} \end{aligned} \tag{34}$$

On the other hand, $\overline{\mu}_{\sigma(K+1-k),n_k} \leq \overline{\mu}_{a_{K+1-k},n_k} \leq \overline{\mu}_{a_{K+1-k},n_k} + \frac{1}{2}\Delta_{(K+1-k)}$. Therefore, using those two observations and (40), we deduce:

$$\begin{aligned} (\mu_{a_j} + \frac{1}{2}\Delta_{(K+1-k)}) - \theta \geq \theta - (\mu_{a_{K+1-k}} + \frac{1}{2}\Delta_{(K+1-k)}), \\ \Delta_{(K+1-k)} \geq 2\theta - \mu_{a_j} - \mu_{a_{K+1-k}} > \theta - \mu_{a_{K+1-k}}. \end{aligned} \tag{35}$$

Thus, we proved that if there is a Type 1 error, then:

$$\Delta_{(K+1-k)} > max(\mu_{a_1} - \theta, \theta - \mu_{a_{K+1-k}}) \tag{36}$$

However, at stage k, only $k-1$ arms have been accepted or rejected, and hence $\Delta_{(K+1-k)} \leq max(\mu_{a_1} - \theta, \theta - \mu_{a_{K+1-k}})$. By contradiction, we conclude that Type 1 error cannot have occurred.

The reasoning process for Type 2 error is similar and omitted for conciseness. This completes the induction and the proof.

□

Proposition 7. *The probability of error of PS satisfies:*

$$e_N \leq 2\alpha K^2 \exp(-\frac{n - \alpha K}{2\alpha \cdot \overline{\log}K \cdot H}) \tag{37}$$

where $H(\alpha) = \max_{i\in\{1,2,\ldots,n\}} i \cdot (|CSV_{\pi_i} - CSV_{\pi_{i+1}}|)^{-2}$, $H = \max_{1\leq j\leq\alpha} H(j)$.

Proof. Consider events ξ_{d_i} for the i-th pre-trained MAB.

$$\xi_{d_i} = \{j \in \{1,2,\ldots,K\}, |\frac{1}{n_k}\sum_{s=1}^{n_k} X_{s,d_1} - f_{j,d_i}| \leq \frac{1}{2}\Delta_{K+1-k}\}$$

Also, consider an event ξ defined as follows.

$$\xi = \{j \in \{1,2,\ldots,K\}, |\frac{1}{n_k}\sum_{s=1}^{n_k} X_s - f_j| \leq \frac{1}{2}\Delta_{K+1-k}\} \tag{38}$$

where f_j is defined as follows.

$$f_j = \frac{\sum_{sim_j\in t,A'\in D}\cos\langle\vec{D},\vec{D'}\rangle \cdot p(t,A')}{\sum_{sim_j\in t,A'\in D}\cos\langle\vec{D},\vec{D'}\rangle} = \frac{\sum_d \cos\langle\vec{D},\vec{d}\rangle \cdot f_{j,d}}{\sum_d \cos\langle\vec{D},\vec{d}\rangle}$$

Letting $w_d = \cos\langle\vec{D},\vec{d}\rangle$, we can rewrite Equation 38 as follows.

$$\xi = \{1 \leq j \leq K, |\frac{1}{n_k}\sum_{s=1}^{n_k}\sum_d w_d \cdot X_{s,d} - \sum_d w_d \cdot f_{j,d}| \leq \frac{1}{2}\sum_{d\in D} w_d\Delta_{K+1-k}\}$$

Suppose Equation 38 is true. Using the absolute value inequality, for any $1 \leq j \leq K$, we have:

$$|\frac{1}{n_k}\sum_{s=1}^{n_k}\sum_d w_d \cdot X_{s,d} - \sum_d w_d \cdot f_{j,d}| \leq \frac{1}{2}\sum_{d\in D} w_d\Delta_{K+1-k}$$

This implies that when $\bar{\xi}$ is true, $\overline{\xi_{d_1}} \cup \ldots \cup \overline{\xi_{d_{|D|}}}$ must be true regardless of w_d. By the law of total probability, we have:

$$P(\overline{\xi_{d_1}} \cup \ldots \cup \overline{\xi_{d_{|D|}}}) \geq P(\overline{\xi_{d_1}} \cup \ldots \cup \overline{\xi_{d_{|D|}}}|\bar{\xi}) \cdot P(\bar{\xi}) = P(\bar{\xi})$$

From the conclusion of Proposition 6, for each $MAB_i \in \{MAB_1, MAB_2, \ldots, MAB_\alpha\}$, we have:

$$P(\overline{\xi_d}) \leq 2K^2 \exp(-\frac{\frac{n}{a} - K}{2\overline{\log}K \cdot H(i)})$$

where $\overline{\log}K = \frac{1}{2} + \sum_{i=2}^{K}\frac{1}{i}$. By union bound, we come to the conclusion that:

$$P(\bar{\xi}) \leq P(\overline{\xi_{d_1}} \cup \ldots \cup \overline{\xi_{d_{|D|}}}) \leq 2aK^2 \exp(-\frac{\frac{n}{a} - K}{2\overline{\log}K \cdot H(a)}) \tag{39}$$

□

REFERENCES

[1] Santiago Ontañón: An Overview of Distance and Similarity Functions for Struc- tured Data. CoRR abs/2002.07420 (2020)

[2] Pei Wang, Weiling Zheng, Jiannan Wang, Jian Pei: Automating Entity Matching Model Development. ICDE 2021: 1296-1307

[3] V. Chvátal and D. Sankoff, "Longest common subsequences of two random sequences," Journal of Applied Probability, vol. 12, no. 2, pp. 306–315, 1975. doi:10.2307/3212444

[4] Vijaymeena M K , Kavitha K , .A Survey on Similarity Measures in Text Min- ing[J].Machine Learning & Applications An International Journal, 2016, 3(1):19- 28.DOI:10.5121/mlaij.2016.3103.

[5] Matthew A. Jaro (1989) Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, Journal of the American Statistical Association, 84:406, 414-420, DOI: 10.1080/01621459.1989.10478785

[6] Dan Tian, Mingchao Li, Yang Shen, Shuai Han: Intelligent mining of safety hazard information from construction documents using semantic similarity and information entropy. Eng. Appl. Artif. Intell. 119: 105742 (2023)

[7] https://github.com/anhaidgroup/deepmatcher/blob/master/Datasets.md

[8] Schulz, Klaus U. and Stoyan Mihov. "Fast string correction with Levenshtein automata." International Journal on Document Analysis and Recognition 5 (2002): 67-85.

[9] Topsoe, F.: Some inequalities for information divergence and related measures of discrimination. IEEE Trans. Inf. Theor. 46(4), 1602–1609 (2000)

[10] Snapper, Ernst . "Metric affine geometry. " Metric Affine Geometry 430.11(1971):1-111.

[11] Saul B. Needleman, Christian D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, Journal of Molecular Biology,Volume 48, Issue 3,1970,Pages 443-453

[12] Shiwei Wei, Yuping Wang, Yiu-ming Cheung, A Branch Elimination-based Efficient Algorithm for Large-scale Multiple Longest Common Subsequence Problem, IEEE Transactions on Knowledge and Data Engineering, (1-1), (2021).

[13] Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho R, Langeland J, Prince V, Wang YL, Westerfield M, Ekker M, Postlethwait JM. 1998. Zebrafish hox clusters and vertebrate genome evolution. Science 282: 1711–1714.

[14] Ballantine, J. P., & Jerbert, A. R. (1952). Distance from a Line, or Plane, to a Poin. The American Mathematical Monthly, 59(4), 242–243. https://doi.org/10.2307/2306514

[15] LEVANDOWSKY, M., WINTER, D. Distance between Sets. Nature 234, 34–35 (1971). https://doi.org/10.1038/234034a0

[16] Reuben R Shamir, Yuval Duchin, Jinyoung Kim, Guillermo Sapiro, Noam Harel bioRxiv 306977; doi: https://doi.org/10.1101/306977

[17] Vijaymeena, M. K.; Kavitha, K. (March 2016). "A Survey on Similarity Measures in Text Mining" (PDF). Machine Learning and Applications. 3 (1): 19–28. doi:10.5121/mlaij.2016.3103.

[18] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. A comparison of string distance metrics for name-matching tasks. In Proceedings of the 2003 International Conference on Information Integration on the Web (IIWEB'03). AAAI Press, 73–78.

[19] https://www.hanlp.com/semantics/dashboard/index

[20] Fagin, R.; Kumar, R.; Sivakumar, D. (2003). Comparing top k lists. SIAM Journal on Discrete Mathematics. 17 (1): 134–160

[21] R. W. Hamming, "Error detecting and error correcting codes," in The Bell System Technical Journal, vol. 29, no. 2, pp. 147-160, April 1950, doi: 10.1002/j.1538-7305.1950.tb00463.x.

[22] Moses S. Charikar. 2002. Similarity estimation techniques from rounding algorithms. In Proceedings of the thiry-fourth annual ACM symposium on Theory of computing (STOC '02). Association for Computing Machinery, New York, NY, USA, 380–388. https://doi.org/10.1145/509907.509965

[23] Multiple Identifications in Multi-Armed Bandits Séebastian Bubeck, Tengyao Wang, Nitin Viswanathan Proceedings of the 30th International Conference on Machine Learning, PMLR 28(1):258-265, 2013.

[24] Radu Cornel Guiasu, Silviu Guiasu: Conditional and Weighted Measures of Ecological Diversity. Int. J. Uncertain. Fuzziness Knowl. Based Syst. 11(3): 283-300 (2003)

[25] Tresoldi, Tiago. (2016). Newer method of string comparison: the Modified Moving Contracting Window Pattern Algorithm.