Sampling-Based Winner Prediction in District-Based Elections

Extended Abstract

Debajyoti Kar
Indian Institute of Science, Bangalore
debajyotikar@iisc.ac.in

Palash Dey
Indian Institute of Technology, Kharagpur
palash.dey@cse.iitkgp.ac.in

Swagato Sanyal
Indian Institute of Technology, Kharagpur
swagato@cse.iitkgp.ac.in

ABSTRACT

In a district-based election, we apply a voting rule r to decide the winners in each district, and a candidate who wins in a maximum number of districts is the winner of the election. We present efficient sampling-based algorithms to predict the winner of such district-based election systems in this paper. When r is plurality (i.e., the candidate receiving a maximum number of votes is declared as the winner) and the margin of victory is known to be at least ε fraction of the total population, we present an algorithm to predict the winner with probability at least $1 - \delta$, whose sample complexity is $O\left(\frac{1}{\varepsilon^4} \log \frac{1}{\varepsilon} \log \frac{1}{\delta}\right)$. We complement this result by proving that any algorithm, from a natural class of algorithms, for predicting the winner in a district-based election when r is plurality, must sample at least $\Omega\left(\frac{1}{\varepsilon^4} \log \frac{1}{\delta}\right)$ votes. We then extend this result to any voting rule r. Loosely speaking, we show that we can predict the winner of a district-based election with an extra overhead of $O\left(\frac{1}{\varepsilon^2} \log \frac{1}{\delta}\right)$ over the sample complexity of predicting the single-district winner under r. We further extend our algorithm for the case when the margin of victory is unknown, but we have only two candidates. We then consider the median voting rule when the set of preferences in each district is single-peaked. We show that the winner of such a district-based election can be predicted with probability at least $1 - \delta$ with $O\left(\frac{1}{\varepsilon^4} \log \frac{1}{\varepsilon} \log \frac{1}{\delta}\right)$ samples.

KEYWORDS

Sampling; Voting; Winner; Prediction; Margin

ACM Reference Format:
Debajyoti Kar, Palash Dey, and Swagato Sanyal. 2023. Sampling-Based Winner Prediction in District-Based Elections: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Voting and elections have always been the de facto method to aggregate different preferences, eventually choosing one of many candidate options. To predict the winner of an upcoming election, a pollster typically samples some votes with the hope that the sampled votes will help him/her correctly predict the winner. However, sampling votes, depending on the sampling requirement and procedure, typically involves substantial cost. Hence, a natural goal of the pollster is to minimize the cost, which often translates to minimizing the number of samples, without compromising the quality (or success rate) of prediction. This same problem is also fundamental in many other applications like social surveys, post election audits [11, 17, 19, 21, 22] etc. Intuitively speaking, this is the winner prediction problem, which is the main focus of our paper.

Bhattacharyya and Dey resolved the sample-complexity of the winner prediction problem for many popular voting rules, for example, k-approval, Borda, approval, maximin, simplified Bucklin, and plurality with run off [3]. We refer to the chapter by Zwicker for an introduction to voting and some common voting rules [23]. In the plurality voting system, each voter votes for one of the candidates and the candidate who receives the maximum number of votes is declared as the winner. We study the winner prediction problem for district-based elections in this paper. In a district-based election, the voters are partitioned into a set of districts. Then some particular voting rule r is used to decide the winner in each district and the candidate winning in the most number of districts wins the election. Indeed, many large-scale elections, for example, Indian general election, US Presidential election, etc., are real-world examples of district-based elections. Since the algorithms in [3] are specific to single-district elections only, they are not applicable in district-based elections. On the other hand, predicting winners in large scale district-based political elections has become norm nowadays. In this work, we fill up this research gap by developing non-trivial algorithms for predicting the winner in district-based elections.

An election is defined by a tuple $(\mathcal{V}, \mathcal{C}, r)$, where $\mathcal{V}$ is a set of N voters, $\mathcal{C}$ a set of m candidates and r is the voting rule (i.e., a function that selects a winner based on the votes of the voters). In a *district-based election*, the set $\mathcal{V}$ of voters is partitioned into k districts, say $(\mathcal{V}_1, \dots, \mathcal{V}_k)$ for some k; each $\mathcal{V}_i, i \in [k]$, is called a district. The overall winner is the candidate who wins in the maximum number of districts. The *Margin Of Victory* (MOV) of an election is defined as the minimum number of votes to be altered to change the winner of the election.

In our sampling model, we are allowed to sample a district from the set of districts uniformly at random. We are also allowed to sample a voter along with her vote uniformly at random from her district. Finally, we are allowed to perform sampling with replacement, that is, the underlying probability space does not change after drawing a sample.

2 OUR CONTRIBUTION

The primary focus of our paper is the (ε, δ)–Winner-Prediction problem, which is defined as follows.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom. © 2023 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

Definition 1 ((ε, δ)–Winner-Prediction). Given an election with N voters partitioned into k districts, and whose margin of victory is at least εN, compute the winner of the election with probability at least $1 - \delta$.

Previous work studied the above problem in the setting of $k = 1$ and an algorithm with optimal sample complexity of $\Theta\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$ is known [3]. Our first result is the following.

Theorem 1. *There is an algorithm for* (ε, δ)–Winner-Prediction *with sample complexity* $O\left(\frac{1}{\varepsilon^4}\log\frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$ *when the plurality voting rule is used to select the winner in each district.*

Our algorithm is fairly simple to state: sample $O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$ districts and from each of these sampled districts, sample $O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon}\right)$ votes and compute their winners using the plurality rule. Finally, output the candidate that wins in maximum number of sampled districts.

We next show that the sample complexity of our algorithm is essentially optimal among a natural class of algorithms.

Theorem 2. *Any algorithm for* (ε, δ)–Winner-Prediction *that works by first sampling* l_1 *districts uniformly at random with replacement and then sampling* l_2 *votes uniformly at random with replacement from each of the sampled districts, must satisfy* $l_1 = \Omega\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$ *and* $l_2 = \Omega\left(\frac{1}{\varepsilon^2}\right)$ *even when there are only* 2 *candidates and all the districts have equal population.*

The above result is principally based on the fact that $\Omega\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$ tosses are necessary to distinguish between two coins whose probabilities of landing up in heads are $\frac{1}{2} + \varepsilon$ and $\frac{1}{2} - \varepsilon$ respectively, with at least $1 - \delta$ probability [1, 5].

We next generalize our result to any arbitrary voting rule r in each district. Let $\chi_r(m, \varepsilon, \delta)$ be the number of samples required so that the predicted winner of a single-district election using rule r with n voters and m candidates, can be made winner by changing at most εn votes. Then, using the prediction algorithm for r, we design an algorithm for (ε, δ)–Winner-Prediction for r with sample complexity $O\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\cdot\chi_r(m, \varepsilon, \varepsilon)\right)$.

In (ε, δ)–Winner-Prediction, we assume that we know some lower bound on the margin of victory of the election. To cater to situations where this information might not be available, we define and study the δ–Winner-Prediction problem.

Definition 2 (δ–Winner-Prediction). Given an election with N voters partitioned into k districts, compute the winner of the election with probability at least $1 - \delta$.

We design algorithms whose sample complexities are a function of the true (unknown) margin of victory of the election.

Theorem 3. *When the plurality rule is used to decide the winners in each district, there is an algorithm for* δ–Winner-Prediction *with sample complexity* $O\left(\frac{1}{\varepsilon^6}\log^2\frac{1}{\varepsilon\delta}\right)$ *when we have only* 2 *candidates, where* εN *is the actual (unknown) margin of victory of the election.*

The above algorithm builds on the observation that there must exist at least $\Omega(\varepsilon^2 k)$ districts in which at least $\frac{1}{2} + \Omega(\varepsilon)$ fraction of voters must have voted for the true winner. Roughly, our algorithm iterates until it reaches a value $\tilde{\varepsilon}$ that satisfies the above property, whereupon it outputs the candidate winning in the majority of (sampled) districts.

Quite often in real-life scenarios the voters are grouped into districts in such a way that each district has roughly the same population. Under this assumption, we are able to improve the sample complexity to $\tilde{O}\left(1/\varepsilon^4\right)$.

Theorem 4. *There is an algorithm for* δ–Winner-Prediction *with sample complexity* $O\left(\frac{1}{\varepsilon^4}\log^2\frac{1}{\varepsilon\delta}\right)$*, where* εN *is the true (unknown) margin of victory, when we have only* 2 *candidates and the number of voters in each district is at most a constant times the average population of a district.*

We next turn our attention towards the median rule. Here there exists an ordering (called the *harmonious ordering*) over the set of candidates and the median with respect to the ordering is declared as the winner of the election. If the harmonious ordering in a district is unknown, we assume that the preference profile of each voter in that district is single-peaked. We show the following result.

Theorem 5. *There exists an algorithm with sample complexity* $O\left(\frac{1}{\varepsilon^4}\log\frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$ *for* (ε, δ)–Winner-Prediction *when the median rule is used to determine the winner of each district.*

3 RELATED WORK

The most immediate predecessor of our (ε, δ)–Winner-Prediction problem is the work of Bhattacharyya and Dey who worked on the same problem but focused only on single district elections [3]. Another classical problem is the winner determination problem in computational social choice. Bartholdi et al. observed that there are popular voting rules, namely the Kemeny voting rule, for which, determining the winner is NP-hard [2]. Hemaspaandra et al. later showed that the above problem is complete for the complexity class $\mathsf{P}_{||}^{\mathsf{NP}}$ [14]. Similar results also hold for the Dodgson and Young voting rules [4, 12, 13, 20].

Our problem is also closely related to the general question: do we need to see all the votes to determine the winner? Conitzer and Sandholm developed preference elicitation policies as a sequence of questions posed to the voters [7]. They showed that finding an effective elicitation policy is NP-hard even for some common voting rules. On the positive side, many effective elicitation policies have been subsequently developed for many important restricted domains and settings [6, 8–10, 15, 16, 18].

4 CONCLUSION

We have initiated the study of the sample complexity for predicting the winner in a district-based election and shown some preliminary results for the problem for some voting rules. Most importantly, we have shown that the sample size remains independent of the number of districts. However, some of our algorithms work only for the case of two candidates and/or when the districts have nearly identical populations. Another research direction is to obtain theoretical guarantees of the algorithm which samples districts without replacement.

REFERENCES

[1] Ziv Bar-Yossef, Ravi Kumar, and D Sivakumar. 2001. Sampling algorithms: lower bounds and applications. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*. 266–275.

[2] John Bartholdi, Craig A Tovey, and Michael A Trick. 1989. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and welfare* 6, 2 (1989), 157–165.

[3] Arnab Bhattacharyya and Palash Dey. 2021. Predicting winner and estimating margin of victory in elections using sampling. *Artificial Intelligence* 296 (2021), 103476.

[4] Felix Brandt, Markus Brill, Edith Hemaspaandra, and Lane A Hemaspaandra. 2015. Bypassing combinatorial protections: Polynomial-time algorithms for single-peaked electorates. *Journal of Artificial Intelligence Research* 53 (2015), 439–496.

[5] Ran Canetti, Guy Even, and Oded Goldreich. 1995. Lower bounds for sampling algorithms for estimating the average. *Inform. Process. Lett.* 53, 1 (1995), 17–25.

[6] Vincent Conitzer. 2009. Eliciting single-peaked preferences using comparison queries. *Journal of Artificial Intelligence Research* 35 (2009), 161–191.

[7] Vincent Conitzer and Tuomas Sandholm. 2002. Vote elicitation: Complexity and strategy-proofness. In *AAAI/IAAI*. 392–397.

[8] Palash Dey and Neeldhara Misra. 2016. Elicitation for Preferences Single Peaked on Trees. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, Subbarao Kambhampati (Ed.). IJCAI/AAAI Press, 215–221. http://www.ijcai.org/Abstract/16/038

[9] Palash Dey and Neeldhara Misra. 2016. Preference Elicitation for Single Crossing Domain. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, Subbarao Kambhampati (Ed.). IJCAI/AAAI Press, 222–228. http://www.ijcai.org/Abstract/16/039

[10] Ning Ding and Fangzhen Lin. 2012. Voting with partial information: Minimal sets of questions to decide an outcome. In *Proceedings of the Fourth International Workshop on Computational Social Choice (COMSOC-2012), Kraków, Poland*.

[11] Joseph Lorenzo Hall, Luke Miratrix, Philip B Stark, Melvin Briones, Elaine Ginnold, Freddie Oakley, Martin Peaden, Gail Pellerin, Tom Stanionis, and Tricia Webber. 2009. Implementing risk-limiting post-election audits in California. In *Electronic Voting Technology Workshop/Workshop on Trustworthy Elections*.

[12] Edith Hemaspaandra, Lane A Hemaspaandra, and Jörg Rothe. 1997. Exact analysis of Dodgson elections: Lewis Carroll's 1876 voting system is complete for parallel access to NP. *Journal of the ACM (JACM)* 44, 6 (1997), 806–825.

[13] Edith Hemaspaandra, Lane A Hemaspaandra, and Jörg Rothe. 2009. Hybrid Elections Broaden Complexity-Theoretic Resistance to Control. *Mathematical Logic Quarterly* 55, 4 (2009), 397–424.

[14] Edith Hemaspaandra, Holger Spakowski, and Jörg Vogel. 2005. The complexity of Kemeny elections. *Theoretical Computer Science* 349, 3 (2005), 382–391.

[15] Tyler Lu and Craig Boutilier. 2011. Robust approximation and incremental elicitation in voting protocols. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

[16] Tyler Lu and Craig Boutilier. 2011. Vote elicitation with probabilistic preference models: Empirical estimation and cost tradeoffs. In *International Conference on Algorithmic Decision Theory*. Springer, 135–149.

[17] Lawrence Norden and Samuelson Law. 2007. *Post-election audits: Restoring trust in elections*. University of California, Berkeley School of Law Boalt Hall.

[18] Joel Oren, Yuval Filmus, and Craig Boutilier. 2013. Efficient vote elicitation under candidate uncertainty. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

[19] Ronald L Rivest and Emily Shen. 2012. A Bayesian Method for Auditing Elections.. In *EVT/WOTE*.

[20] Jörg Rothe, Holger Spakowski, and Jörg Vogel. 2003. Exact complexity of the winner problem for Young elections. *Theory of Computing Systems* 36, 4 (2003), 375–386.

[21] Philip B Stark. 2008. Conservative statistical post-election audits. *The Annals of Applied Statistics* 2, 2 (2008), 550–581.

[22] Scott Wolchok, Eric Wustrow, J Alex Halderman, Hari K Prasad, Arun Kankipati, Sai Krishna Sakhamuri, Vasavya Yagati, and Rop Gonggrijp. 2010. Security analysis of India's electronic voting machines. In *Proc. 17th ACM Conference on Computer and Communications Security*. ACM, 1–14.

[23] William S. Zwicker. 2016. Introduction to the Theory of Voting. In *Handbook of Computational Social Choice*, Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia (Eds.). Cambridge University Press, 23–56. https://doi.org/10.1017/CBO9781107446984.003