Authorship Attribution via KNN, DWKNN, and GRNN (October 2018)

Sarp Aykent, Jordan Cox, and Blake Schilleci

*Abstract*—Authorship Attribution is available through multiple methods. The methods included in this paper are “K-Nearest Neighbor”, “K-Nearest Neighbor (Distance Weighted)”, and a General Regression Neural Network. The KNN algorithm uses the shortest distance from the query instance to the training element to then predict the output. DWKNN is very similar to KNN but instead uses a distance-weighted function to give the closest neighbors higher priority (or weighted more heavily). A GRNN uses a set amount of hidden functions (fire strengths) to assist the query instance in predicting the desired output. This paper will outline the process of creating these algorithms to implement our data set.

# INTRODUCTION

Authorship Attribution is done in many different types of data sets and can range widely in size [30]. In our experiment, we use a rather small data set when dealing with only 20+ writing samples. Authorship Attribution can be done with high accuracy rates and high reliability. Our data set contains 5 different authors and this presents us with a small data set. With this being said, covering the same event week to week(i.e. Mississippi State football game), it begins to get a little muddy when allowing the program to predict authors. Optimizing algorithms that have stood the test of time in order to best suit our data set can be tedious but extremely clear. Understanding that, each week all five authors will most definitely include the same key players and more specifically, the same statistics.

Through our genetic algorithms, we were able to optimize each method and create an accuracy rate that would show to be reliable. After trial and error, and working through error rates as high as 92%, the roadblocks began to reveal themselves. Understanding how each method works through the data set was imperative in gaining a lower error rate. Repeatability was the main theme in fixing our issues. Eliminating repeated articles and same authors that are about the same games finally allowed for our programs to run with high accuracy.

# Methodology

For our DWKNN algorithm, the baseline was unsurprisingly inaccurate due to our data set being so messy and arbitrary. The best way to optimize the algorithm was to remove any unnecessary articles that related to the same games. Since the author we're querying for doesn't have an article for the same game, this gave an unfair advantage to the other authors with articles of the same week. Another optimization we used in DWKNN is giving each author a weight, using a Gaussian distribution. This gave the program a higher success rate in order to find the correct author. Understanding how to optimize this algorithm was honestly very simple. In a general sense, a clearer picture was painted so that there was an obvious choice for each training element.

For KNN, the optimization was slightly similar but deals with maximum distances for the k-neighborhood. This limits the distance required to go since there is a boundary on the lengths at which the query must go to predict the outcome. Accuracy was abysmal to begin with, but after optimization it was improved by approximately 10% when k=3 and k=5.

With the GRNN, a genetic algorithm is created to optimize the direct outputs of the training set for the lowest possible error percentage. This algorithm will create random individuals based off of a Gaussian distribution and will use these individuals to update the desired output vector. It performs this action so that the success rate will be increased.

# Experiment

We wrote our genetic algorithms in Python, and we gave it two datasets to traverse, our CASIS-25 dataset and our SEC Sports Writer Dataset (Mississippi State)[1 2 3 4 5 6 7 8 9 10 11 12 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29]. All three algorithms take in and read each data set to perform whichever method. Each algorithm has been optimized in its own way and improvements have been made. This is exciting, but the optimizations still leave room for error.

For DWKNN, the CASIS-25 dataset is read in first using the baseline version of DWKNN. The success rates will be recorded. Next, the Sports Writer dataset will be read and ran through the baseline DWKNN with success rates being outputted. Finally, the Sports Writer dataset will be ran through the optimized version of DWKNN to show the improvements in success rates [...].

For KNN, the same process is done as it is in DWKNN. CASIS-25 is the beginning dataset to be ran through the baseline version KNN, with error and success rates being outputted. After this, the Sports Writer dataset will be put through the baseline KNN and the error/success rates will be presented. Lastly, the Sports Writer dataset will go through the optimized KNN and the improved success rates will be shown [...].

With the GRNN, the genetic algorithm (as mentioned above) will only work on the Sports Writer dataset. Once again, CASIS-25 will be ran through the GRNN and the success rates of predicted output will be created. Then the Sports Writer dataset will be read and ran by the GRNN, this is the baseline version. To finish off this portion of the experiment, the GRNN will once again read in the Sports Writer dataset but it will run it through the optimized version of the GRNN which will create better success rates [...] With the optimized version of the GRNN, the results do vary. In the best case scenario, the error will be 0%, but in the worst case it will be the baseline percentages. This is a display of inconsistency with the GRNN. Naturally, this will happen when creating individuals with a random number from a Gaussian distribution.

Our results are consistent with KNN and DWKNN, but the inconsistency does leave the door open to different possibilities when optimizing the GRNN.

# Results

DWKNN:

CASIS-25

K = 1 => error: 50.0  - success 50.0

K = 3 => error: 48.0  - success 52.0

K = 5 => error: 44.0  - success 56.00000000000001

K = 99 => error: 53.0  - success 47.0

Sports Writer (baseline)

K = 1 => error: 53.84615384615385  - success: 46.15384615384615

K = 3 => error: 53.84615384615385  - success: 46.15384615384615

K = 5 => error: 65.38461538461539  - success: 34.61538461538461

K = 25 => error: 65.38461538461539  - success: 34.61538461538461

Sports Writer (optimized)

K = 1 => error: 7.6923076923076925  - success: 92.3076923076923

K = 3 => error: 19.230769230769234  - success: 80.76923076923077

K = 5 => error: 23.076923076923077  - success: 76.92307692307692

K = 21 => error: 15.38461538461538 - success: 84.61538461538461

KNN:

CASIS-25

K = 1 =>  error: 50.0%, success: 50.0%

K = 3 =>  error: 55.0%, success: 44.999999999999996%

K = 5 =>  error: 54.0%, success: 45.999999999999996%

K = 99 =>  error: 100.0%, success: 0.0%

Sports Writer (baseline)

K = 1 =>  error: 53.84615384615384%,

success: 46.153846153846156%

K = 3 =>  error: 69.23076923076923%,

success: 30.76923076923077%

K = 5 =>  error: 76.92307692307693%,

success: 23.076923076923073%

K = 25 =>  error: 100.0%,

success: 0.0%

Sports Writer (optimized)

K = 1 =>  error: 53.84615384615384%,

success: 46.153846153846156%

K = 3 =>  error: 61.53846153846154%,

success: 38.46153846153846%

K = 5 =>  error: 65.38461538461539%,

success: 34.615384615384615%

K = 25 =>  error: 61.53846153846154%,

success: 38.46153846153846%

GRNN:

CASIS-25 (sigma = 0.2)

K = 1 => error: 50.0%, success: 50.0%

K = 3 => error: 48.0%, success: 52.0%

K = 5 => error: 45.0%, success: 55.0%

K = n => error: 46.0%, success: 54.0%

Sports Writer (baseline; sigma = 0.03)

K = 1 => error: 53.84615384615385%, success: 46.1538461538%

K = 3 => error: 53.84615384615385%, success: 46.1538461538%

K = 5 => error: 53.84615384615385%, success: 46.1538461538%

K = n => error: 53.84615384615385%, success: 46.1538461538%

Sports Writer (optimized; 20 offspring; sigma = 0.18)

K = 1 => (average after 5 generations) error: 0.0 - 10.0%,

success: 90.0 - 100.0%

K = 3 => (average after 5 generations) error: 3.0 - 20.0%,

success: 80.0 - 97.0%

K = 5 => (average after 5 generations) error: 7.0 - 26.0%,

success: 74.0 - 93.0%

K = n => (average) error: 11.5384%, success: 88.4616%

# Breakdown of Work

**Sarp Aykent:** Programmed DWKNN Algorithm and Optimized DWKNN. Collected data related to success rates of DWKNN.

**Jordan Cox:** Programmed GRNN Algorithm and Optimized GRNN. Assisted in writing paper and revising. Collected data related to success rates of GRNN

**Blake Schilleci:** Programed KNN Algorithm and Optimized KNN. Wrote the paper and collected references and articles for the Sports Writer dataset. Collected data related to success rates of KNN.

References

[1] A. Press. “No. 18 Mississippi State tops SFA 63-6 in Moorhead's debut.” *ESPN,* 1 Sept. 2018. [Online]. www.espn.com.

[2] A. Press. “Fitzgerald leads no. 18 miss st to 31-10 rout of K-State.” *ESPN,* 8 Sept. 2018. [Online]. www.espn.com.

[3] A. Press. “No. 18 Mississippi State breezes by SF Austin.” *SEC Sports,* 8 Sept. 2018. [Online]. www.secsports.com.

[4] A. Press. “No. 18 Mississippi State stifles Kansas State 31-10.” *SEC Sports,* 8 Sept. 2018. [Online]. www.secsports.com.

[5] H. Cloud. “Defensive dog collars, slam dunks, are all part of a culture change.” *The Reflector,*12 Sept. 2018. [Online]. www.reflector-online.com.

[6] H. Cloud. “Bulldog offense lifts off, defense smothers in first win for Moorhead.” *The Reflector,*2 Sept. 2018. [Online]. www.reflector-online.com.

[7] J. Coleman. “Bulldogs becoming disruptive on defense.” *Bulldogs Becoming Disruptive on Defense | Starkville Daily News*, 2 Sept. 2018. [Online]. Starkvilledailynews.com.

[8] T. Horka. “Running dack Kylin Hill leads Mississippi State to 31-10 victory over Kansas State.” *The Clarion Ledger*, The Clarion-Ledger, 8 Sept. 2018. [Online]. www.clarionledger.com.

[9] T. Horka. “Thompson, Hill lead Mississippi State to 63-6 stomping of Stephen F. Austin.” *The Clarion Ledger*, The Clarion-Ledger, 2 Sept. 2018. [Online]. www.clarionledger.com.

[10] B. Hudson. “MSU's defense shows versatility in dominating Kansas State.” *The Commercial Dispatch*, 10 Sept. 2018. [Online]. www.cdispatch.com.

[11] B Hudson. “Columbus' Hill has rousing start to season.” *The Commercial Dispatch*, 2 Sept. 2018. [Online]. www.cdispatch.com.

[12] D. Skretta. “Thrill from Hill: sophomore running back produces in MSU victory.” *Bulldogs Becoming Disruptive on Defense | Starkville Daily News*, 8 Sept. 2018. [Online]. starkvilledailynews.com.

[13] S. Nirkhi, R. V. Dharaskar, and V. M. Thakare, “An experimental study on authorship identification for cyber forensics,” *Research Gate* Accessed on: Oct 15, 2018. [Online]. Available: https://www.researchgate.net/publication/289433295\_An\_Experimental\_Study\_on\_Authorship\_Identification\_for\_Cyber\_Forensics

[14] D. Brandt. “Fitzgerald, No. 16 Mississippi State Rout La.-Layayette.” *ESPN*, ESPN Internet Ventures, 16 Sept. 2018, www.espn.com/college-football/recap?gameId=401012278.

[15] D. Brandt. “Snell Leads Kentucky Past No. 14 Mississippi State 28-7.” *ESPN*, ESPN Internet Ventures, 23 Sept. 2018, www.espn.com/college-football/recap?gameId=401012286.

[16] D. Brandt. “Mullen, Florida Beat No. 23 Mississippi State 13-6.” *ESPN*, ESPN Internet Ventures, 30 Sept. 2018, www.espn.com/college-football/recap?gameId=401012298.

[17] D. Brandt. “Fitzgerald Leads Miss St to 23-9 Win over No. 8 Auburn.” *ESPN*, ESPN Internet Ventures, 7 Oct. 2018, www.espn.com/college-football/recap?gameId=401012304.

[18] J. Coleman. “A-Train Gets Back on Track for No. 14 MSU.” *A-Train Gets Back on Track for No. 14 MSU | Starkville Daily News*, 16 Sept. 2018, starkvilledailynews.com/node/124495.

[19] G. Graves. “Bad Night: Wildcats Put Damper on Plans of Bulldogs.” *Bad Night: Wildcats Put Damper on Plans of Bulldogs | Starkville Daily News*, 23 Sept. 2018, starkvilledailynews.com/node/124552.

[20] J. Coleman. “Searching for Answers: Bulldogs Left Scratching Their Heads as Former Coach Leaves with a Win.” *Searching for Answers: Bulldogs Left Scratching Their Heads as Former Coach Leaves with a Win | Starkville Daily News*, 30 Sept. 2018, starkvilledailynews.com/node/124604.

[21] J. Coleman. “Sweet Relief: MSU Finds End Zone, Victory over Auburn.” *Sweet Relief: MSU Finds End Zone, Victory over Auburn | Starkville Daily News*, 7 Oct. 2018, starkvilledailynews.com/node/124650.

[22] B. Hudson. “Fitzgerald Looks Dominant as No. 16 Bulldogs Slice and Dice Ragin' Cajuns.” *The Commercial Dispatch*, www.cdispatch.com/sports/article.asp?aid=68524.

[23] B. Hudson. “Kentucky Upsets No. 14 MSU.” *The Commercial Dispatch*, www.cdispatch.com/sports/article.asp?aid=68661.

[24] B. Hudson. “MSU's Offense Struggles in Loss to Florida.” *The Commercial Dispatch*,

www.cdispatch.com/sports/article.asp?aid=68856.

[25] B. Hudson. “Running Games Propels MSU Past No. 8 Auburn.” *The Commercial Dispatch*,

www.cdispatch.com/sports/article.asp?aid=69007.

[26] H. Cloud. “Wilcox Honored, MSU Rolls to Victory in Third Game of Season.” *The Reflector*, 19

Sept. 2018, www.reflector-online.com/sports/article\_c849b1bc-b977-11e8-9f39-3faf1e8637ad.html.

[27] H. Cloud. “Bulldogs Knock down No. 9 Tigers.” *The Reflector*, 8 Oct. 2018, www.reflector-

online.com/sports/article\_bf4c2cc4-c9ef-11e8-8bb1-8b906830601d.html.

[28] H. Cloud. “Moving on from Kentucky, MSU Focused on First Home SEC Game.” *The Reflector*, 27

Sept. 2018, www.reflector-online.com/sports/article\_a15ade6a-c25e-11e8-bb32-ab7b74da8d2a.html.

[29]H. Cloud. “Mullen's Return Proves Unhappy for Bulldogs.” *The Reflector*, 30 Sept. 2018, www.reflector-online.com/sports/article\_36296bcc-c469-11e8-af8e-eb779ea01df7.html.

[30] Luyckx, K. and Daelemans, W. (2018). *Authorship Attribution and Verification with Many Authors and Limited Data*. [online] Aclweb.org. Available at: http://www.aclweb.org/anthology/C08-1065 [Accessed 18 Oct. 2018].