A Supervised classification model (kNN) to Predict Infectious Disease Characteristics in Bees

Yeshwanth Kallati
Department of Computer Science and
Engineering
University of California, Riverside
yk001@ucr.edu

ABSTRACT

In this paper, we describe a a supervised classification model (kNN) which predicts infectious disease characteristics in Bees. Bees are essential pollinators of major food crops. But their numbers are decreasing a lot day by day. Bees interact with other bees and they travel from a hive to hive. If a diseased bee mates with another healthy or diseased bee the offspring will be affected. So, studying the flight characteristics helps us to classify between healthy bees and infected bees, using which we can be able to stop the spreading of diseases and helps to increase the population of bees.

We propose a k Nearest Neighbors (kNN) approach, a supervised classification model which is based on Euclidean distance metric for classification of infectious bees. We formulate the algorithm as an optimization task which learns to distinguish between control group — which are healthy group — and experimental group — which is exposed to nosema virus—. To the best of our knowledge, we implemented the model with a z-score normalized dataset. We have used the z-score normalization technique for the problem, achieving approximately 69% accuracy over the records in the dataset. kNN elucidates the flight patterns of bees corresponding to Radio-frequency identification (rfid) tags. Finally, the project aims to provide a simple supervised tool yielding high accuracy and also making it easy-to-use for the practitioners.

Keywords

kNN, Computational Entomology, Bees, Z-score Normalization.

1.INTRODUCTION

Bees are an integral part in pollinating the food crops. Recent changes in the environment are bringing down the bee population. With the data set created by the Center for Integrative Bee Research at UCR, we have access to the mating flight behavior of healthy and infected males carrying RFID tags. We are able to study the flight characteristics such as frequency, time, between healthy and infected males. A Fungal pathogen-Nosema Ceranae is introduced into the bees in the experimental group. Weather conditions play a key role in mating flight behaviors of bees. There is much data collected on bees, such as pollination flights patterns and mating flight patterns. However, a lot of the data is stored and is not in use mostly. With the rise in trends of data science and analytics of big data sets, the idea is to be able to use that data to gain a better understanding of the bees. There is a much needed observation that should be studied from the mating flight behaviors of bees. Bees control a lot of portion of pollinating work of food crops.

Not only the nosema are the cause for the overall reduction of population, but pesticides and the changes in the environment are also the factors. The problem with the prediction of the diseases in bees is data collection. It is a very hard process to collect the data from bees and often times the collected data is not accurate and a lot of preprocessing is to be done to obtain the pure form of data which can be used for the training of the model. Another problem is that most of the times the data is corrupted. Some bees can go missing and might not represent the true reading through the rfid tags. The problem sometimes get even worse where the disease has no effect on the overall flight behavior of the bees. To overcome these problems and to achieve a good quality in the model we transform the data using various techniques. One of them is normalizing the data using the Z-score normalization technique. The Z-score normalization is a technique where we subtract the mean from every value of the dataset and then divide it by the standard deviation. This way the data is more normalized and can be better analyzed. We develop a supervised model and perform the cross validation to ensure the overall quality of the data is prominent. We implement the kNN with the Euclidean distance metric on a real-world normalized dataset. Our methodologies scale to large real-world scenarios and can be employed in a variety of applications.

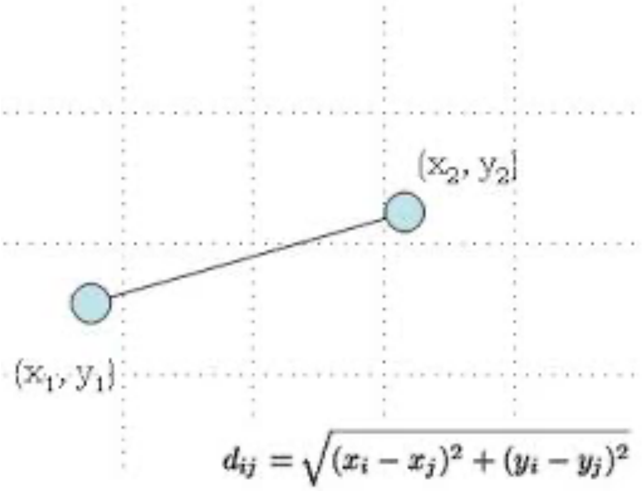
2. BACKGROUND AND RELATED WORK

Nosema ceranae in drone honey bees (Apis mellifera) [1] The article expresses that Nosema apis, a microsporidian pathogen of bumble bees, was the main microsporidian pathogen to have tainted bumble bees until an another pathogen was found called Nosema Ceranae. The spread of Nosema Apis for the most part happens an oral-fecal course inside the hives however the spread of Nosema ceranae is as yet obscure. The objective of the paper was to decide whether Nosema ceranae normally taints bumble bees. Their techniques were to arbitrarily test honey bees from provinces situated in apiaries and concentrate their genomic DNA and perform quantitative polymerase chain response. The honey bees were tested over various months. For the outcomes, they broke down automaton honey bees by sorting them in three gatherings. The gatherings were called Drone Pupae, In-Hive Drones, and Flying Drones. Generally, They found that automaton honey bees had low degrees of normal contamination of Nosema ceranae. They found that each gathering had a lot of being normally tainted with Nosema ceranae. Automaton Pupae had low contamination is the long periods of May and June. The creators propose that disease in ramble pupae might be because of brood nourishment or defilement. In-Hive rambles were increasingly tainted around the period of June. By and large, Flying automatons had the most significant level of Nosema and might be because of floating to different hives in the apiary. Floating is where the honey bees come back to an inappropriate hive because of perplexity.

KNN model-based approach in classification : This article underscores on the downsides of the customary K-Nearest Neighbors (KNN) and an approach to conquer those disadvantages . The creators start by clarifying how KNN being a lethargic student needs unique characterization for a huge storehouse. KNN's reliance of K is talked about and it's predisposition by k is featured. They proposed another kNN type characterization strategy called kNNModel, which consequently computes K. Their proposed strategy was demonstrated to be ideal as far as order exactness. This KNN model for the information lessens the reliance on K, making the arrangement quicker. Wang H proposes a strategy which takes a gander at various arrangements of closest neighbors instead of only one to discover K. Alongside the Euclidean separation they depend on number of information focuses inside a neighborhood area and the similitude between the most distant point in a locale to it's middle. The K esteem is dictated by the quantity of datapoints in a district. This methodology depends on the help estimation of different arrangements of closest neighbors to all the more likely know the class of an information point. This calculation makes an inductive information model from the preparation informational index and depends on that model for order. The value of this model is that it has a progressively exact arrangement as it is less subject to K. one con is that it's moderately moderate, and requires O(n^2) to order another occasion. Stretching out from their methodology, an extra spotlight can be put on characterization of unrepresented information to diminish time unpredictability.

3. PROPOSED METHOD

This project is an attempt to add to the related work. We will look at an informational collection made by the Center for Integrative Bee Research that began at the University of Western Australia and now at UCR, where the mating flight conduct has been examined in sound and tainted guys conveying RFID labels. The exploratory gathering was tainted with a contagious pathogen known as Nosema ceranae. The test enabled us to examine flight exercises of guys for a mind-blowing duration and enables us to consider different distinctive flight attributes, for example, recurrence, time, term, which can likewise be looked at among solid and contaminated guys. The information will likewise incorporate climate information that will be scratched from the web. This is verifiable information that contains distinctive climate conditions every day for the time allotment that the honey bee informational collection incorporates. We propose a k Nearest Neighbors (kNN) approach, an supervised classification model which depends on Euclidean distance metric for order of irresistible honey bees. We figure the calculation as an enhancement task which figures out how to recognize control group — which are healthy group — and experimental group — which is presented to nosema infection—. As far as we could possibly know, we actualized the model with a z-score standardized dataset. We have utilized the z-score standardization method for the issue, accomplishing around 69% precision over the records in the dataset. kNN clarifies the flight examples of honey bees comparing to Radio-recurrence ID (rfid) labels. At last, the task means to give a straightforward managed instrument yielding high exactness and furthermore making it simple to-use for the professionals.We will apply characterization and prescient investigation models to the information, in order to uncover a few examples or differentiations that could be utilized in future, enormous scale contemplates. We use the z-score normalization on the dataset to achieve a better accuracy. We use the kNN with the eculidean distance metric which is shown in the figure below.



We use this distance metric and calculate the distance and get the top k nearest neighbors which are points in the dataset. The algorithm is shown below. The algorithm is implemented from scratch and achieves an accuracy of 69% with an f1 score of 0.77. This algorithm also serves as a way for us to classify different classes of bees. For example it can differentiate between Control (Healthy) and experimental (infected) groups. This model is written in python from scratch and developed using Jupyter Notebooks.

Algorithm

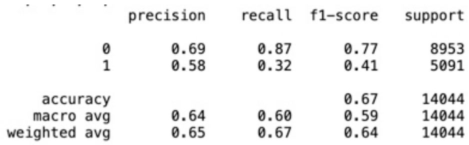
```
def k_nearest_neighbors(data, predict, k):
    if len(data) >= k:
        warnings.warn('K is set to a value less than total
voting groups!')

    distances = []
    for group in data:
        for features in data[group]:
            euclidean_distance =
np.linalg.norm(np.array(features)-np.array(predict))
            distances.append([euclidean_distance,group])

    votes = [i[1] for i in sorted(distances)[:k]]
    vote_result = Counter(votes).most_common(1)[0]
[0]
    return vote_result
```

4. EXPERIMENTAL EVALUATION

In this section we will present the experimental aspects and the results of the study that was performed to investigate and evaluate the classification of bees. We maintain a different group of bees called the experimental group. Those groups of bees are affected with the virus and they are labelled as infected. We use those labels to determine the accuracy of our model .we calculate the confused matrix, precision, recall, f1 score, and accuracy of the model. We use the following code to populate the lists with prediction values.

```
actual_list = []
pred_list = []
for group in test_set:
    for data in test_set[group]:
        vote = k_nearest_neighbors(train_set, data, 81)
        actual_list.append(group)
        pred_list.append(vote)
        if group == vote:
            correct += 1
        total += 1
```

So, we populate the actual and prediction lists and pass it to the classification_report function and get the following results.



```
              precision    recall  f1-score   support

           0       0.69      0.87      0.77      8953
           1       0.58      0.32      0.41      5091

    accuracy                           0.67     14044
   macro avg       0.64      0.60      0.59     14044
weighted avg       0.65      0.67      0.64     14044
```

We use the cross validation and separate the train and test data. We have divided the data into three parts where the first part which is 70% is for training and the 20% is for validation and the rest of the 10% is for validation. Overall our model performed well with an accuracy of 68%. We were able to achieve a precision of 0.69, recall of 0.87 and the f1-score of 0.77.

5. DISCUSSIONS AND CONCLUSIONS

The results of our model are good given the quality of the dataset. The data which is available is still not accurate and the results only reflect that. Further the project can be improved by introducing more complex algorithms and also improving the quality of the data. Also, neural networks can be used to better understand the relationship of various attributes and the patterns in the data.

6. ACKNOWLEDGMENTS

Our thanks to UCR - Center for Integrated Bee Research for allowing us to use and modify the data they had collected.

7.REFERENCES

1. Liu, W., & Chawla, S. (2011, May). Class confidence weighted knn algorithms for imbalanced data sets. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 345-356). Springer, Berlin, Heidelberg.
2. Traver, Brenna E., and Richard D. Fell. "Nosema ceranae in drone honey bees (Apis mellifera)." *Journal of invertebrate pathology* 107.3 (2011): 234-236.
3. Su Jinshu, Zhang Bofeng, Xu Xin, “Advances in Machine Learning Based Text Categorization”, Journal of Software, Vol.17, No.9, 2006, pp.18481859
4. Ma Jinna, “Study on Categorization Algorithm of Chinese Text”, Dissertation of Master’s Degree, University of Shanghai for Science and Technology, 2006
5. Wang Jianhui, Wang Hongwei, Shen Zhan, Hu Yunfa, “A Simple and Efficient Algorithm to Classify a Large Scale of Texts”, Journal of Computer Research and Development, Vol.42, No.1, 2005, pp.8593
6. Li Ying, Zhang Xiaohui, Wang Huayong, Chang Guiran, “V ector-Combination-Applied KNN Method for Chinese Text Categorization”, Mini- Micro Systems, Vol.25, No.6, 2004, pp.993996
7. Wang Yi, Bai Shi, Wang Zhang’ou, “A Fast KNN Algorithm Applied to Web Text Categorization”, Journal of The China Society for Scientific and Technical Information, Vol.26, No.1, 2007, pp.6064
8. Fabrizio Sebastiani, “Machine learning in automated text categorization”, ACM Computer Survey, Vol.34, No.1, 2002, pp. 1-47
9. Conger., S., and Loch, K.D. (eds.). Ethics and computer use. Commun. ACM 38, 12 (entire issue).
10. Mackay, W.E. Ethics, lies and videotape... in Proceedings of CHI '95 (Denver CO, May 1995), ACM Press, 138-145.
11. Schwartz, M., and Task Force on Bias-Free Language. Guidelines for Bias-Free Writing. Indiana University Press, Bloomington IN, 1995.
12. ACM SIG PROCEEDINGS template. http://www.acm.org/sigs/pubs/proceed/template.html.
13. Belur V, Dasarathy, “Nearest Neighbor (NN) Norms, kNN Pattern Classification Techniques”, Mc Graw-Hill Computer Science Series, IEEE Computer Society Press, Las Alamitos, California, 1991,pp.217-224
14. Yang Lihua , Dai Qi, Guo Yanjun, “Study on KNN Text Categorization Algorithm”, Micro Computer Information, No.21, 2006, pp.269271
15. Wang Yu, Wang Zhengguo, „A fast knn algorithm for text categorization”, Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007, pp.3436-3441
16. http://www.nlp.org.cn/docs/doclist.php?cat_id=16 [15] Yang Y, Pedersen J O, “A comparative study on feature selection in text categorization”, ICNL,1997,pp.412-420
17. Xinhao Wang, Dingsheng Luo, Xihong Wu,Huisheng Chi, “Improving Chinese Text Categorization by Outlier Learning”, Proceeding ofNLP-KE'05pp. 602-607
18. Jin Yang, Zuo Wanli, “A Clustering Algorithm Using Dynamic Nearest Neighbors Selection Model”, Chinese Journal of Computers, Vol.30, No.5, 2005, pp.759762
19. Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” in Lecture Notes in Computer Science 904, P. M. B. Vita ́nyi, Ed. Berlin: Springer, 1995, pp. 23–37.
20. A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” Journal of the Royal Statistics Society -B, vol. 39, no. 1, pp. 1–38, 1977.
21. J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, California: Morgan Kaufmann, 1993.
22. S. T. Dumais, J. Platt, D. Heckerman, and M. Sahami, “Inductive learning algorithms and representation for text categorization,” in Proc. of the 7th ACM International Conference on Information and Knowledge Management (CIKM’98), Bethesda, MD, 1998, pp. 148–155.
23. Y. Freund and L. Mason, “The alternating decision tree learning al- gorithm,” in Proc. of the 16th International Conference on Machine Learning (ICML’99), Bled, Slovenia, 1999, pp. 124–133.
24. R. E. Schapire and Y. Singer, “Improved boosting algorithms using confidence-rated predictions,” in Proc. of the 11th Annual Conference on Computational Learning Theory (COLT’98), New York, 1998, pp. 80–91.
25. M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, “Learning multi-label scene classification,” Pattern Recognition, vol. 37, no. 9, pp. 1757–1771, 2004.
26. G. Salton, “Developments in automatic text retrieval,” Science, vol. 253, pp. 974–980, 1991.
27. P. Pavlidis, J. Weston, J. Cai, and W. N. Grundy, “Combining microarray expression data and phylogenetic profiles to learn functional categories using support vector machines,” in Proceedings of the 5th Annual International Conference on Computational Biology, Montre ́al, Canada, 2001, pp. 242–248.