West Nile Virus Prediction Using Machine Classification

Frank Hiemstra
Dept of Electrical and Computer Engineering
University of Virginia
Charlottesville, VA 22903
fhh7ph@virginia.edu

Andrew Norton
Dept of Computer Science
University of Virginia
Charlottesville, VA 22903
apn4za@virginia.edu

Abstract

Since 2002 Chicago has seen an outbreak of the West Nile virus, a virus primarily spread to humans through disease-ridden mosquitoes. The severity of this epidemic is as serious as the consequences of being infected with the virus, which include but are not limited to a persistent fever, serious neurological illnesses, or even death. The Chicago Department of Public Health (CDPH) has established a comprehensive surveillance and control program to combat the spread of the virus. This program involves surveying selected areas of the city, where samples of mosquitoes tested for the virus. Next, the city then sprays infected areas to eliminate these virus-carrying mosquitoes. Not all mosquitoes in Chicago carry the West Nile virus, so targeting which areas of the city have infected mosquitoes, and at what time, is an important task. Weather conditions, testing data from the CDPH program, and spraying data are all used in this project to predict when and where different species of mosquitoes will test positive for the West Nile Virus. The ability to predict this is crucial to not only city of Chicago residents, but also people around the world, as people come to the city for business and tourism. This project has the capability to give strategic, statistical-based direction for which areas in Chicago and at what time the CDPH should spray chemicals to kill mosquitoes. The task will be accomplished using regression and classification analysis.

1 Target task

From an early age, children in primary school are told they can make a difference. They are told good deeds can create a snowball effect. At first a few people are affected, then a community, a region, a state, a country, the globe. This is a fantastic encouragement we should be instilling in the youth; however, the opposite can also be true. A negative chain reaction is what this project aims to prevent. Chicago has been infected with West Nile virus, of which the first outbreaks recorded were in Israel in the 1950s. West Nile virus can cause physical and neurological harm, but even death. Since the first sightings of the virus in Chicago in 2002, the 2.72 million Chicago residents have been very concerned. However, infecting the city has much larger ramifications than harming only its citys residents. Chicago estimated it had 50.2 million people from around the globe tour the city in 2014. This magnifies West Nile virus from a city concern to a serious global threat. People visiting the city could become contaminated with the virus in Chicago and bring it back to their prospective homeland.

In 2004 the Chicago Department of Public Health (CDPH) initiated a program to fight the virus. The program includes testing mosquitoes in various sites around Chicago for the virus, as well as spraying combative chemicals to eliminate these disease-carrying insects. Given this data from the past, along with NOAA weather data this project aims to give the CDPH a model to predict at what time different areas in Chicago will have mosquitoes infected with the virus. This project aims to yield results that will stop the spread of West Nile virus from mosquitoes to people living in or visiting the city of Chicago by predicting the location and times of occurrences of West Nile virus in mosquitoes.

2 Previous work

The National Center for Atmospheric Research in Boulder, Colorado, released a report earlier this year associating meteorological conditions and incidence of West Nile virus in the United States [Hahn, *et al.*]. Without using machine learning, they looked at how variations in temperature or precipitation alter the likelihood of increased cases of West Nile virus. As a study with nation-wide scope, this "broad strokes" approach is useful for creating rough estimates of mosquito activity over large time intervals. However, applying machine learning to this problem could improve the specificity of their predictions.

In a more location-targeted study, the Geographic Information Systems and Spatial Analysis Laboratory at University of Illinois at Urbana-Champaign attempted to predict the spread of West Nile virus in the Chicago metropolitan area [Ruiz, *et al.*]. In their 2010 paper, they used linear regression, regression trees, and random forests to create a statistical model of the mosquito infection rate per week over the 2004-2008 timeframe. From 2011-2014, they used this model to predict mosquito infection rate. Earlier this year, Kaggle ran a competition focused on West Nile virus prediction in Chicago, based on a similar (but more current) dataset.

3 Relevance to machine learning

This problem, as we will approach it, is inherently rooted in machine learning. We aim to predict when and where mosquitoes carrying the West Nile virus will appear in Chicago based on historical data. While one could approach this topic using more "pure" mathematical modelling or a simplified statistical model, machine learning will allow us to utilize far more data and create much more precise predictions.

4 Proposed method

The task at hand is supervised learning; in particular, we are facing a classification problem; given the weather history and disease test results for a set of areas, classify each area as containing or not containing West Nile Virus. As such, our first approach will be to use a Support Vector Classifier.

Since we will be classifying our samples into only two distinct classes, we will use simple, binary classification techniques. As stated previously, we will use a Support Vector Classifier (Vapnik, 1995). Due to its simplicity, we will start with a Linear Support Vector Classifier. The performance of this classifier will be compared with SVCs with other kernels using k-fold cross validation.

Regardless of how the SVC performs, our team will use an ensemble approach in classifying our samples. As mentioned before, we will investigate other Support Vector Classifiers in addition to the linear kernel. To take full advantage of a mixed ensemble, though, we will consider completely different classifiers. In particular, the two other classifiers we will explore are the K-Nearest-Neighbors and Decision Tree approaches. By using the ensemble method, we can With the ensemble approach, we will base our decision on how to classify a given sample based on the average of the different classifiers; however, we may weight the influence a specific classifier has differently.

In classification contests over the past several years, Decision Tree-based approaches have performed substantively well, but (due to the model complexity) have a tendency to overfit data. Because of the limited number of actual cases of West Nile Virus in our training data (see Evaluation Metrics section), there is a strong risk of overfitting. For this reason, we expect to use this model as simply a member of an ensemble instead of the "primary" classifier we are considering.

For this problem, our features will include, but are not limited to: date, location, humidity, pressure, outside temperature, wind speed, past spraying history, past testing history. After initial analysis and testing, our team will consider weighting the features when using them for classification.

This task will require a fair amount of preprocessing. First, some of our features contain categorical data. This means we will need to digitally discretize these features. For the continuous parameters, we will need to use the preprocessing technique of scaling to avoid numerical difficulties during calculations. As a first attempt, we will linearly scale, both the training and testing data, to fit a range of [0,1]. This is beneficial because it decreases the time it takes for gradient descent and other numerical methods to converge, as well as preventing certain features from being inherently weighted simply due to the size of the numbers involved. Better scaling may be achieved by centering the features so they have mean 0.5 and then dividing by the variance of the sample.

Lastly, there are missing values some of our feature data, which means we will need to do preprocessing to assign a value to each. Our first attempt at this will be imputation using nearest neighbor; however, our team would like to explore a more complex EM based approached to this preprocessing. If neither of the above work, we will substitute the missing value with the mean value of the feature.

5 Experimental design and data

We will be using a dataset from the Kaggle competition mentioned under *Previous Work* [Kaggle]. Much of the data is given in a time-based format, recording the results of an sampling performed at a particular time. It is important to note, here, that we should only use historical data for prediction; that is, if we are attempting to predict West Nile virus presence on the 5th of June, we should not rely on weather reports following June 5.

The data may be divided into three categories, which are in separate files:

1. Weather data
2. Spray location and times
3. Mosquito sampling results

The weather data consists of standard NOAA measurements: temperature, dewpoint, precipitation, windspeed, and a variety of other lesser-known statistics. It is organized in a "daily" format, where each row of the set is one weather station's measurements for a given day.

The "spray location and times" table is simply a log of the date, time, and latitude/longitude of each occurrence of insecticide spraying the city of Chicago performed.

The last data table contains results from trap-and-test samples performed across Chicago. Each row contains the test results from a single location, on a single day, of a single mosquito species. The exact location is given in the form of an address, but these have been converted to longitude/latitude coordinates (within a given accuracy) using Geocoder. The "results" of a test are the number of mosquitoes captured and the number of these mosquitoes that tested positive for West Nile virus.

6 Evaluation Metrics and Model Selection Design

Of the ten-thousand training data samples we have been given, a few more than five hundred contain West Nile virus. The skewed nature of the data set will make it somewhat difficult to train classifiers. That said, when considering the previous work done on a problem such as this, our team would deem ourselves as successful to achieve a testing accuracy of 65%. Seeing that the city of Chicago is over 200 square miles, even a prediction with this accuracy would of far more practical use than simply guessing.

The difference between recording a high accuracy versus a low accuracy can be in model selection. One objective is to steer clear of overfitting. Example kernels for our SVC are a linear, radial basis function (RBF), and polynomial kernels. We will use the k-folds cross validation technique on the training data to choose hyperparameters, then apply the best kernel with hyperparameters to the testing data.

In the same vein, we will use k-fold cross validation to help select the right number of neighbors to consider in the K-Nearest-Neighbors classifier. Because the fewer neighbors considered increases the "complexity" of the model *and* our dataset is quite skewed, selecting the K value is important.

If we choose too low of a K value, we will tend to overfit our data, but if K is too high, then we will not be able to predict with accuracy.

For completeness sake, our team will investigate some principles in the literature regarding kernel selection, such as diffusion kernel, Fisher kernel, string kernel, etc.

7 Experimental Results

7.1 3-Fold Cross Validation

Model	Accuracy	Std. Deviation
Support Vector Classifier	0.5642	0.366
Naive Bayes	0.2109	0.140
Random Forest	0.6704	0.405
Decision Tree	0.6234	0.395
5-Nearest Neighbor	0.7346	0.327

7.2 Precision, Recall, and F1-Score

7.3 Testing Accuracy

After tuning our model parameters with cross validation (and the F1-Score), we submitted our testing output to the Kaggle submission server. We receive two results: a “public” score and a “private” score (this is a feature of Kaggle competitions, which allows closed-source programs to be judged on a separate dataset than open-source programs). The table below summarizes our results for the private scoreboard:

Model	Accuracy
Support Vector Classifier	0.646
Naive Bayes	0.535
Random Forest	0.513
Decision Tree	0.530
5-Nearest Neighbor	0.5022

8 Conclusion and Discussions

9 Team justification

We have a biological issue at hand with mosquitoes in the city of Chicago carrying the West Nile virus. A large step in solving this problem will come from analyzing data in order to make an adaptive, predictive solution. This task will be accomplished through an intense effort involving statistics, mathematics, and computer programming. Our team consists of two members with different, complimentary backgrounds that meet and exceed the required skills needed to solve this problem. Andrew is a Computer Science and Mathematics double major, whereas Frank is an Electrical Engineering graduate student. Not only are the technical skills met, our team understands the scope and the significance of the issue in order to get to the root problem. The presence of the West Nile virus in Chicago carries implications much further than a city concern, but rather a world issue. Our team will treat this project as such.

In order to come up with a solution to this problem, many solutions should be examined. Our team has a wide breadth of knowledge of various classification techniques which will allow us to investigate different solutions. In a classroom setting we have been exposed to expository teaching on classification subject matter, while being given real world problems to solve. This team is motivated by helping others, and is excited to use their college education to help the city of Chicago. Furthermore, one of our team has had exposure to working with weather data and analyzing trends in both an academic research environment, and in industry. The intelligence of our group is met with great curiosity and vigor to accomplish this task.

This project holds significance for the second author, Andrew Norton, because one could apply the results of this project to protecting the residents of Chicago. As a third year undergraduate, he is still deciding on the focus of study for his future research; however, he certainly wants to see his efforts to advance a field protect mankind from the dangers of the world.

References

Hahn, M., Monaghan, A., Hayden, M., Eisen, R., Delorey, M., Lindsey, N., . . . Fischer, M. (2015). Meteorological Conditions Associated with Increased Incidence of West Nile Virus Disease in the United States, 2004-2012. *American Journal of Tropical Medicine and Hygiene*, 1013-1022.

Ruiz, M., Chaves, L., Hamer, G., Sun, T., Brown, W., Walker, E., . . . Kitron, U. (n.d.). Local impact of temperature and precipitation on West Nile virus infection in Culex species mosquitoes in northeast Illinois, USA. *Parasites Vectors Parasites & Vectors*, 19-19.

West Nile Virus Prediction. (n.d.). Retrieved October 12, 2015, from https://www.kaggle.com/c/predict-west-nile-virus

Vapnik, Vladimir, and Corinna Cortes. “Support-Vector Networks.” Machine Learning 20 (1995): 273-97. http://csee.wvu.edu/x̃inl/library/papers/comp/ML/svm.pdf.