www.ioirp.com
International Journal of Innovative Research in Computer Science and Engineering (IJIRCSE)
ISSN: 2394-6364, Volume – 1, Issue – 5. April 2015

OUTBREAK PREDICTION OF BIRD FLU VIA SATELLITE TRACKING

[1]R.Trinita Vaiz, [2]S.Sangeetha
[12]Computer Science and Engineering, Panimalar Engineering College, Chennai, India
[1]trinivaiz@yahoo.co.in, [2]sangics185@gmail.com

***Abstract*— Advanced satellite tracking technologies have collected huge amounts of wild bird migration data. Biologists use these data to understand dynamic migration patterns, study correlations between habitats, and predict global spreading trends of avian influenza. The research discussed here transforms the biological problem into a machine learning problem by converting wild bird migratory paths into graphs. H5N1 outbreak prediction is achieved by discovering weighted closed cliques from the graphs using the mining algorithm High-weight closed clique mining (HELEN). The learning algorithm HELEN-p then predicts potential H5N1 outbreaks at habitats. This prediction method is more accurate than traditional methods used on a migration dataset obtained through a real satellite bird-tracking system. Empirical analysis shows that H5N1 spreads in a manner of high-weight closed cliques and frequent cliques.**

Keywords*— Outbreak, manifold based, chronology, pathogenic, correlation

I. INTRODUCTION

Avian bird flu has been an ongoing topic of international concern. Here, we transform the bird-migration data analysis problem into a high-weight closed clique mining problem, and we propose a novel, High weight closed clique mining (HELEN) algorithm, which our prediction algorithm HELEN-p then uses for accurate H5N1 outbreak prediction. The H5N1 virus outbreaks in poultry in 2003, 2004, and 2009 had an unprecedented geographical impact in Asia. The H5N1 virus is a highly pathogenic avian influenza that has emerged in southern China in the mid-1990s. A large number of wild birds died as a result of the highly pathogenic virus in Qinghai Lake, China, in 2005. The number of protected bar-headed geese had decreased 5 to 10 percent worldwide due to the epizootic disease, as estimated in 2009.However, effective tracking systems and data analysis tools have been lacking for a long time in China. The study on the relationship between the spread of the H5N1 virus and the bird-migration network wasn't conducted on a large scale. This situation is greatly improved now. The GPS devices continuously transmitted tracking signals to the satellite and the US Geological Survey processing unit distributed the data to researchers. Biologists found that bird migration routes in a small area are best viewed as graph. Patterns like cliques rather than simple location sequences on a small scale. It's therefore important to understand the role that migratory birds play in the ecology and transmission patterns of H5N1 by integrating data on habitats, seasonal movement chronology, routes, dates, and locations of H5N1 outbreak events. Recently, several studies have shown that H5N1 viruses in Qinghai Lake spread with the bird migration Patterns .Most of these analyses were conducted at a relatively coarse level of granularity and the methods for discovering the correlations of bird migration routes have limited predictive power. Here, we mine the bird-movement pattern data and learn the relationship between graphical clique patterns and virus propagation. In particular, we use vertex weights to evaluate the seriousness of H5N1 virus transmission. Weights are differently defined by using the degree of a habitat or vertex, the time that birds stay at a certain habitat, or the density of the birds in a particular habitat. These weighted graph features can make the virus prediction model more accurate because we can use them to better estimate the correlations among the habitats. As a result, our prediction algorithm HELEN-p can help accurately predict future H5N1 outbreak from the migration graphs. In our previous work, we analyzed bird virus outbreaks via mining bird migration data such as sequence rule and subgraph mining. In this article, we focus on how to predict future possible bird virus outbreak locations with machine learning methods. Specifically, our prediction method is based on mined high-weight closed cliques, some newly developed habitat correlation criteria, and two machine learning algorithms More importantly, with LapRLS (Laplacian-based regularized least square)we generalized the idea of label propagation in manifold based, semi supervised learning to H5N1spreads in the bird migration network.

Fig 1. A GPS tracking device attached to a bird. Ecologists captured and attached devices to the birds to monitor and analyze their migration.

II. HELEN ALGORITHM

We first introduce the basic concepts and principles of weighted graphs, and then describe the high weight closed clique mining and H5N1 virus outbreak prediction algorithms.

In our graph-based model, a bird habitat is denoted by a node (vertex) and a migration route is denoted by an edge. A clique *C* is a graph with fully connected edges. If a graph *G* contains a clique *C*, then *G* is said to be a support graph of *C*.

A. FREQUENCY SUPPORT:

The frequency-support of a clique *C* is defined as the ratio of the number of support graphs over the total number of graphs in a database *D*,

$$\sup port_f(c) = \frac{\sum_{G \in D} I(C \subseteq G)}{|D|} \quad \text{..... (1)}$$

Where $\sum_{G \in D} I(C \subseteq G)$ is the number of support graphs of clique *C*, and |*D*| is the number of graphs in the database.

Given a support threshold *θf*, a clique *C* is a frequent clique if support*f*(*C*) = *θf*. In addition, if there doesn't exist another clique *C*′ satisfying *C*⊆*C*′and support*f*(*C*′)= support*f*(*C*), *C* is a frequent closed clique. Closed cliques are important since they greatly reduce the number of child cliques with the same support level. Frequent-closed clique mining finds all frequent closed cliques from a graph database. The weight of a vertex *v* is denoted by weight(*v*). We consider three weighting ideas in this work:

- ***W*frequency,** which measures how frequently a bird flies among different habitats.
- ***W*time =*t*arrive-*t*leave**, which measures how long a bird stays at a certain habitat, where *t*arrive and *t*leave are the bird's arrival and departure Times.
- ***W*density**, which measures bird density in the habitat, and is calculated by using the habitat's area size divided by the number of migration records received by the satellite tracking system from that habitat. The weight of a graph *G* is given by **weight (*G*) =Σ*v*∈*G*weight (*v*).**

B. WEIGHT SUPPORT

The weight-support of a clique *C* is defined as

$$\sup port_w(c) = \frac{weight(C)\sum_{G \in D} I(C \subseteq G)}{\sum_{G \in D} weight(G)} \quad \text{..... (2)}$$

Where the numerator $\sum G \in D\ I(C \subseteq G)$ denotes the total weight of the clique *C* in database *D*, and the denominator $\sum G \in D$ weight(G) is simply a normalization term. Given a support threshold *θw*, a clique *C* is a high-weight support clique if support*w*(*C*) >*θw*. In addition, if no other clique *C*′ exists that satisfies *C* ⊆ *C*′and support*w*(*C*′) ≥support*w*(*C*), then *C* is a high-

weight support closed clique (HWCC). We wish to find all frequent and closed cliques from graph database *D* with respect to the vertex weight.

C. GRAPH WEIGHT SUPPORT

The graph-weight support of a clique *C* is defined as

$$\sup port_g(C) = \frac{\sum_{G \in D} I(C \subseteq G) weight(G)}{\sum_{G \in D} weight(G)} \quad \text{.....(3)}$$

Where the numerator $\sum_{G \in D} I(C \subseteq G)$ weight(G) denotes the total weight of support graphs of clique *C* in database *D*, and the denominator $\sum G \in D$ weight *G* is again for normalization. Given a support threshold *θg*, a clique *C* is a high-graph-weight-support clique if support*g*(*C*) ≥*θg*. In addition, if there doesn't exist a clique *C′* satisfying *C* ⊆ *C′*and support*g*(*C′*) =support*g*(*C*), *C* is a high-graph-weight support closed clique (HGWCC).The downward-closure property, which has been widely used to accelerate pattern-mining algorithms, states that any child pattern of a frequent pattern is also frequent. Hence, if no *k*-1-patterns are frequent, we don't need to explore *k*-patterns. However, we observe that the downward-closure property doesn't hold true in HWCC mining. This also causes difficulties for mining algorithms. If it can be proved that if any k-1-clique *C*[*k-1*] isn't a high-graph-weight-support clique, then *k*-clique *C*[*k*] isn't either. This downward-closure property is useful in the process of enumerating cliques. If we know that a *k* -1-clique *C*[*k*-1] isn't a high-graph-weight-support clique, there's no need to enumerate any *k*-clique. It can be also proved that if *θw* =*θg*, then HWCC ⊆ HGWCC.

D. CALCULATING HABITAT CORRELATION

Our prediction method also involves two types of habitat correlations: Location and clique-based. For any two habitats *i* and *j*, the location-based correlation is defined by the distance *dij* of the two habitats, calculated using

$$\frac{1/d_{ij}}{\max_{ij_{ij}} 1/d_{ij}} \quad \text{..... (4)}$$

Where the denominator, max*ij* 1/*dij*, is a normalization term to make the correlation in the range of [0, 1].We consider two types of distance in our correlation estimation:

- The Euclidean distance $d_{ij}^{ec} = \sqrt{(\varphi_i - \varphi_j)^2 + (\lambda_i - \lambda_j)^2}$, where (ϕ_i, λ_i) and (ϕ_j, λ_j) are the latitude and longitude of habitats *i* and *j*,respectively.
- The great-circle distance $d_{ij}^{gc} = r\Delta\sigma_{ij}$, where *r* is the radius $\Delta\lambda = \lambda_i - \lambda_j$,and

$$\Delta\sigma_{ij} = \arctan\times \frac{\sqrt{(\cos\phi_j \sin\Delta\lambda)^2 + (\cos\varphi_i \sin\phi_j - \sin\varphi_i \cos\phi_j \cos\Delta\lambda)^2}}{\sin\phi_i \sin\phi_j + \cos\phi_i \cos\phi_j \cos\Delta\lambda} \quad \text{..... (5)}$$

For any two habitats *i* and *j*, the clique-based correlation is defined by using the weighted supports of closed cliques to which *i* and *j* belong:

$$c_{ij}^{w} = \frac{\sum_{C \varepsilon C} I((i,j) \subseteq C \sup port^{w}(C)}{\max_{ij} \sum_{C \in C} I((i,j \in C) \sup port^{w}(C)} \quad \text{..... (6)}$$

Where C is a set of HWCCs, and $\sum_{C \in C} I((i, j) \subseteq C)$

support$w(C)$denotes the summation of the weighted support of the closed cliques to which the habitats *i* and *j* belong.

III. PREDICTION ALGORITHM

We take the following pseudo codes in the prediction of H5N1 virus outbreaks:
Input: Graph database *D*, vertex weight, threshold θg and θw, positive instance p, number of predicted habitats *k*;
Output: A ranked list of *k* predicted habitats.

1. Call the HELEN algorithm to obtain HWCC.
2. Calculate the correlations of any two habitats according to Equations 3 or 4 using the mined HWCC.
3. Run the *k*NN or LapRLS algorithm to find the top *k* likely outbreak habitats.

The problem setting of our prediction task is transductive learning rather than inductive learning, where the input includes one positive instance that is, labeled training data many unlabeled instances that is, unlabeled test data and correlations among the instances. A common supervised machine learning method trains a prediction model using labeled training data only, for which one single positive instance isn't sufficient. The two machine learning methods *k*NN and LapRLS are explained as follows. We hypothesize that a H5N1 outbreak is highly correlated with the migration network, which is reflected in the mined high-weight closed cliques. We verify this hypothesis in the experimental section later in the article. Given a habitat with an H5N1 outbreak and the habitat correlation, we can rank the remaining habitats and obtain the top *k* habitats with the largest correlation based on the *k*NN method. We denote the corresponding HELEN-p variant as HELEN-p(*k*NN).Under a kernel-learning approach, we take the originating habitat of the H5N1 outbreak as a single positive instance. We predict other outbreak habitats by using the LapRLS method, where the normalized Laplacian matrix *L* is calculated based on a habitat correlation

$$W = [C_{ij}^{w}] \in R^{n*n}$$
$$L = I - D^{-1/2} W D^{-1/2} \quad \ldots\ldots\ldots\ldots\ldots\ldots (7)$$

Where $D = diag(W1)$

Where **I** is an identity matrix and **1** is the vector with all entry values of 1.Then, we apply the LapRLS objective function with a single positive instance

$$\min f'Lf + \frac{a}{n} \| f - y \|_F^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (8)$$

Where $f \in R^{n*1}$ is the prediction vector,y is the label vector with y=(1 if i=p,0 if $i \neq p$),$\|.\|_F$ Frobenius norm, and *a* is the tradeoff parameter. Hence, the final obtained score vector ***f*** can be used to rank the remaining habitats and find the top *k* habitats with the highest probability of an H5N1 outbreak. We denote the corresponding HELEN-p variant as HELEN-p (LapRLS).Compared with the HELEN-p(*k*NN) method, HELEN-p(LapRLS) has the potential of bridging two habitats beyond *k*-nearest neighbors, because it can propagate the label via local connections, which is also supported by our experimental results.

A. DATA COLLECTION

Ecologists randomly captured 59 birds from different flocks and tied a battery-powered GPS device to each of them. They collected nearly 1 million migration records of the 59 birds by 25 December 2009.We selected 29 bar-headed geese for our subsequent analysis of the same type of birds. The 29 bar-headed geese correspond to 29 graphs in our algorithms, and each graph contains the same 103 nodes corresponding to 103 habitats. We used the reverse transcription-polymerase chain reaction technique to confirm whether a bird is or isn't infected with the virus, and hence to determine the prevalence of H5N1 in Qinghai Lake. We tested 1,055 samples .The experiments confirmed that 12 bar- headed geese, three ruddy shelducks, and 14 brown-headed gulls are positive for an H5N1 subtype. These data are compared to the total numbers of birds of the three types and it can be seen that the prevalence of H5N1 in Qinghai Lake was high. To obtain the relationship between migratory birds and H5N1 outbreaks, they extracted information about H5N1 outbreaks from the Ministry of Agriculture of the People's Republic of China

Database and the World Organization for Animal Health (OIE) Database for the period of February 2004 to May 2009.We conduct empirical studies of H5N1 outbreak analysis and prediction using the mined cliques in the following two subsections. We applied the HELEN algorithm to those 29 graphs to extract cliques. If we only consider its frequency support (support*f*= 3/29), *C*15 would be pruned. However, this clique has a weight of 0.13, 0.16, and 0.052, respectively, according to *W*frequency,*W*time, and *W*density weighting strategies, and it contributes to more than 5.2 percent of the total time of the birds' spring migration time.

For example, while birds prefer to stay at habitat 4 (H4), three cases of H5N1 outbreak are reported. In addition, this clique shows that the habitat H4 has a strong correlation with its neighboring habitats (H1, H2,H3, and H5) under the high weight of *W*density. Interestingly, habitats (H2,H3, and H5) are also reported to have H5N1 outbreaks. The weight of those habitats does reflect the possibility of virus transmission.

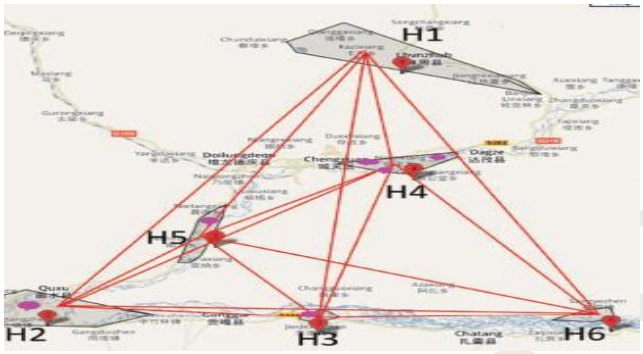

Fig : 2

A total of 24 percent of our mined cliques have a low frequency but a high weighted support. This magnifies the importance of weight clique mining, because otherwise, these low frequency cliques would be pruned by the traditional frequent-closed clique mining algorithms. We can see two important points:
the approach of using clique-based correlation is much better than that of using the habitats geometric information, confirming the usefulness of the bird satellite tracking system or migration network in habitat correlation estimation; and although the clique-based correlation might fail to build connections between two habitats that never appear in any of the same cliques, as shown by the results of HELEN-p(*k*NN),HELEN-p(LapRLS) can complement this weakness via label propagation.

Habitat	$w_{frequency}$	w_{time}	$w_{density}$	**Outbreak cases**
H1	18	100	800	N/A
H2	34	140	130	1
H3	35	109	103	1
H4	31	173	270	3
H5	48	9	69	1
H6	24	19	78	N/A

Table: 1 detailed information of the habitats and weight of clique

Acknowledgment

We can see that using a relatively larger threshold improves prediction performance in most cases. This observation can be explained by the fact that a reduction of noise in the clique weights can result in better correlation estimation. However, using a too-large threshold could reduce the prediction performance, which makes sense because the correlation between two habitats might not appear when using too few selected closed cliques. Therefore, we can conclude that using a relatively higher threshold is better in prediction, which supports our assumption that H5N1 spreads via high-weight closed cliques. In this article, we've developed a novel H5N1 outbreak prediction algorithm (HELEN-p) that makes use of the mined cliques and machine learning methods. Our assumption that H5N1 spreads via high-weight closed cliques and frequent cliques is also supported by our experimental results. We'll explore more sophisticated algorithms to integrate different weighting strategies and contextual constraints.

References

[1] H. Chenetal.,"Avian Fu: H5N1 Virus Outbreak in Migratory Waterfowl," *Nature*, vol. 436, July 2005, pp. 191–192.

[2] J. Liu et al., "Highly Pathogenic H5N1Influenza Virus Infection in Migratory Birds," *Science*, vol. 309, no. 5738, 2005,p. 1206.

[3] M. Tang et al., "Exploring the WildBirds' Migration Data for the Disease Spread Study of H5N1: A Clustering and Association Approach," *Knowledge and Information Systems*, vol. 27, May 2011,pp. 227–251.

[4] Z. Kou et al., "The Survey pf H5N1 Flu Virus in Wild Birds in 14 Provinces ofChina from 2004 to 2007," *PLOS ONE*, vol. 4, no. 9, 2009; doi:10.1371/journal.pone.0006926.

[5] Y.-S. Hou et al., "Distribution and Diversity of Waterfowl Population in Qinghai Lake National Nature Reserve," *Acta Zootaxonomica Sinica*, vol. 34, no. 1,2009, pp. 184–187.

[6] M. Tang et al., "Birds Bring Flues? Mining Frequent and High Weighted Cliques from Birds Migration Networks,"*Proc. 15th Int'l Conf. Database Systems for Advanced Applications*, vol. 2, 2010, pp. 359–369.t

[7] M. Belkin P. Niyogi, and V. Sindhwani,"Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," *J. Machine Learning Research*, vol. 7, Nov. 2006, pp. 2399–2434.

[8] T. Vincenty, "Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations," *Survey Review*, vol. 23, no. 176, 1975, pp. 88–93.

[9] X. Ling et al., "Spectral Domain-Transfer Learning," *Proc. 14th Ann. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, 2008, pp. 488–496.

[10] Q. Yang, "A Theory of Conflict Resolution in Planning," *Artificial Intelligence*, vol. 58, nos. 1–3, 1992, pp. 361–392.