An Innovative Big Data Predictive Analytics Framework over Hybrid Big Data Sources with an Application for Disease Analytics

Joglas Souza¹, Carson K. Leung¹(✉) and Alfredo Cuzzocrea²

¹ Department of Computer Science, University of Manitoba, Winnipeg, MB, Canada
kleung@cs.umanitoba.ca
² iDEA Lab, University of Calabria, Rende, Italy

Abstract. Nowadays, big data are everywhere. Examples of big data include weather data, web-search data, disease reports, as well as epidemic data and statistics. These big data can be easily generated and collected from a wide variety of data sources. A data science framework—such as predictive analytics framework—helps mining data from various big data sources to find useful information and discover knowledge, which can then be transformed into wisdom for appropriate actions. In this paper, we present an innovative big data predictive analytics framework over hybrid big data sources. To demonstrate the effectiveness and practicality of our framework, we conduct several case studies, including one on applying the framework to disease analytics. More specifically, we integrate, incorporate and analyze weather data and web-search data to predict and forecast dengue cases based on a hybrid of three kernels in support vector machine (SVM) ensemble. Results show how our predictive analytics framework benefits health agencies in disease control and prevention.

1 Introduction

Nowadays, big data are everywhere. Examples of big data include social network data [16], meteorological records [4], omic data (e.g., genomic data) [20], web-search data, disease reports, as well as epidemic data and statistics. These big data can be easily generated and collected from a wide variety of data sources. A data science solution [8]—such as predictive analytics framework [1,7]—helps mining data from various big data sources to find useful information and discover knowledge, which can then be transformed into wisdom for appropriate actions. In this paper, we present an innovative big data predictive analytics framework over hybrid big data sources. Here, our predictive analytics model uses a hybrid of three different kernels (i.e., mathematical functions) in support vector machine (SVM):
- linear kernel,
- polynomial kernel, and
- *Radial basis function* (RBF) kernel.

To demonstrate the effectiveness and practicality of our big data predictive analytics framework, we conduct a case study on applying the framework to disease analytics. Accurate and efficient prediction of disease based on a wide variety of data from various sources helps prevention and control of diseases such as viral diseases. Examples of viral diseases include:

- Dengue fever, caused by dengue virus (DENV)
- Severe acute respiratory syndrome (SARS), with outbreak in 2002–2004, caused by SARS coronavirus (SARS-CoV)
- Avian influenza (aka *bird flu*), with global spread in 2004–2007, mainly caused by influenza A virus of subtype H5N1, denoted as A(H5N1) or simply H5N1.
- 2009 H1N1 influenza pandemic (aka swine flu), mainly caused by influenza A virus of subtype H1N1, denoted as A(H1N1) or simply H1N1.
- Middle East respiratory syndrome (MERS), with outbreak in 2012–2015, caused by MERS coronavirus (MERS-CoV)
- Coronavirus disease 2019 (COVID-19), with outbreak in 2019–2020, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)—aka 2019 novel coronavirus (2019-nCoV)

As a preview, in this paper, we use the framework to analyze weather data and web-search data to predict and forecast dengue cases as it is listed as a potential threat among 10 diseases for 2019 by the World Health Organization (WHO)\(^1\).

Dengue is a mosquito-borne viral disease. According to the WHO\(^2\), dengue virus has been transmitted by female mosquitoes—mainly of the species *Aedes aegypti* (aka yellow fever mosquito) and *Aedes albopictus* (aka *Stegomyia albopicta*, Asian tiger mosquito, or forest mosquito)—in recent years. Their incubation period ranges from 4 to 10 days. These mosquitoes are also vectors of other viruses such as chikungunya, yellow fever and Zika. Dengue is widespread throughout the tropics, with local variations in risk influenced by rainfall, temperature, relative humidity, and unplanned rapid urbanization. It has outbreaks in more than 100 countries worldwide, and is estimated to infect 50–100 million people yearly. As infected humans are source of contamination for non-infected mosquitoes, the spread of the disease can be very fast. The first symptoms appear 2–7 days after the mosquito bite. There are four dengue virus (DENV-1, DENV-2, DENV-3, and DENV-4) with different severities in which symptoms range from severe headaches, fever, muscle and joint pains until plasma leaking, severe bleeding or organ damaged that can lead to death. A person that was infected by one of the virus gets immunized only for the specific one. There are no vaccinations available for dengue. The only way to stop the spread of the disease is preventing the procreation of the mosquito. Among the methods for prevention are the proper covering of water containers or any other material that

---

1. [https://www.who.int/health-topics/dengue-and-severe-dengue](https://www.who.int/health-topics/dengue-and-severe-dengue).
2. [https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue](https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue).
accumulates water as the female mosquito needs water in order to lay its eggs. Also, during outbreaks the population can wear repellents, clothing that covers skin and install screens at windows and doors. The procreation occurs under some weather conditions (e.g., rainfall, temperature, relative humidity) and the population density.

Due to the relation between weather conditions and the proliferation of the dengue vector mosquito, statistical and machine-learning models can be used for forecasting the number of dengue cases based on weather conditions [11,12,21]. The detection of such outbreaks in advance would give to public health agencies time for dealing with the high number of cases and even decrease them. Moreover, web-search of terms also related to diseases in platforms like Google correlates with the outbreaks [3,17]. Hence, in this paper, our key contributions include:

- design and evaluation of an innovative big data predictive analytics framework over a wide variety of data sources, and
- application of our framework for disease analytics.

In particular, our framework uses a hybrid of three kernels—namely, linear, polynomial, and RBF kernels—in SVM ensemble to predict and forecast dengue cases. It combines both (i) weather variables and (ii) the popularity of web-search terms related to dengue, mines and analyzes these data. The data were collected between the years of 2004 and 2008 from San Juan, Puerto Rico, where dengue endemic and outbreaks still occur in 1994, 1998, 2007 and 2010 according to the Center for Disease Control and Prevention (CDC)\footnote{https://www.cdc.gov/dengue/}.

The remainder of this paper is organized as follows. The next section describes related works. Section 3 presents our innovative big data predictive analytics framework over hybrid big data sources and its application for disease analytics. Section 4 shows results of evaluation on our framework for analyzing weather data and web-search terms to predict the number of dengue cases in San Juan, Puerto Rico. Finally, conclusions are drawn in Sect. 5.

## 2 Related Works

Hii et al. [11] only used weather variables (namely, temperature and rainfall) as inputs to a Poisson multivariate regression model. The data were collected in Singapore between the years of 2002 and 2011. Results showed an accuracy of 96% distinguishing outbreaks from non-outbreaks between 2002 and 2010 and 98% accuracy for 2011. In their study, they also observed a lag time between the weather conditions and the development of the mosquito. This lag comprehends the time that takes for the egg hatching, larvae period and lastly the mosquito born. High incidences of dengue cases in 9–16 weeks after some weather events in their findings are consequences of cumulative rainfalls, average high temperatures and humidity.

In contrast, Kesorn et al. [13] observed that only weather variables would not be sufficient for precisely forecasting dengue outbreaks. They introduced
“mosquito infection rate” as a new feature, and used six different machine-
learning models for comparison in two different scenarios (namely, with vs. with-
out the new feature). Their results on the data collected in Thailand showed that
the included variable on average improved the accuracy by 4.32% at a price of
more computation and longer runtime, due to the difficulty in collecting and
analyzing the infection rate. Their chosen model—namely, RBF-kernel SVM—
considering the new variable as one of the features led to an accuracy of 96.29%.

Yusof and Mustaffa [21] observed that neural network (NN) and least square
SVM (LS-SVM) require simpler calculations and less complexity in training than
the regular SVM. Their comparison on applying the NN and LS-SVM to data
collected from five Malaysian cities showed that the LS-SVM outperformed the
NN in accuracy and runtime. Specifically, the LS-SVM led to an average accuracy
of 86.74% and runtime of 0.14 s (cf. 65.58% accuracy and 452.94 s for the NN).

Similarly, Husin et al. [12] compared the NN with a nonlinear regression
model (NLRM). Their comparison on applying the NN and NLRM to the same
data (i.e., the same five Malaysian cities as Yusof and Mustaffa) showed that the
NN led to smaller mean square errors (MSE): 0.028 for the NN and 26.054 for
the NLRM. Note that MSE measures the quality of the estimator based on the
errors. The closer the value to zero, the better is the model.

Guo et al. [10] tested model performance in six different machine-learning
algorithms on data collected from five provinces in China. They incorporated
both (i) weather data and (ii) the search query index from Baidu (a Chinese
web-search tool) into their model, which uses support vector regression (SVR)
with linear kernel. Specifically, they collected—on a weekly basis—the frequency
of terms related to “dengue” searched by the population in the studied provinces.
Their model was shown to lead to an average accuracy of 98% among the
five provinces. However, they did not measure the extra improvement in pre-
diction accuracy brought by the incorporation of these two additional features
(i.e., weather data and search query index from Baidu).

Milinovich et al. [17] used a more extensive approach for infectious disease
surveillance based entirely on web-search terms for predicting outbreaks. Specif-
ically, they used Google as the web-search tool for querying the frequency of the
terms. They observed that people tend to search on the Internet for treatments
and symptoms of a particular disease once they get sick. Hence, high frequency
of a term (e.g., “dengue”) could be an indication of an epidemic. Back in 2005
and 2006, there were a high frequency of the search term “bird flu” due to the
global spread of bird flu, especially in China and Turkey. Models for predicting
influenza and dengue were created based on the frequency of terms. Such an
idea has been incorporated into platforms such as (i) Google Flu Trends and
(ii) Google Dengue Trends⁴, which are no longer publishing current estimates
of flu or dengue fever based on search patterns. Despite the positive impacts of
their approach, they also observed that external factors (e.g., media influence,
language barriers, false positives) may affect the number of searches not corre-
ling with actual cases of diseases. For example, media can drive such scenarios

⁴ https://www.google.org/flutrends/about/.
depending on what they publish. As another example, the same terms can be used to describe a variety of different situations with no correlation to diseases. Hence, as their models would need to be validated in each specific scenario before being used by health agencies, the models were not yet ready as replacement of existing methods used by health agencies.

Chan et al. [3] applied a web-search model for predicting dengue cases in Bolivia, Brazil, India, Indonesia and Singapore. Their model presents high correlations from 0.82 to 0.99 between the picks of dengue searches on Google and the epidemics. However, they observed that false alarms may happen due to some factors already presented. For example, for people go online searching for dengue, the model assumes the people have already contracted dengue (instead of suffering from other disease with similar symptoms). Moreover, there is possibility that people may misdiagnose their symptoms. Depending on the studied regions, there may be insufficient data due to limitations on Internet access by the population in the studied countries. This explains their model was suggested as a complement (instead of replacement) for traditional approaches.

3 Our Predictive Analytics Framework

3.1 Overview

Let us give an overview of our innovative big data predictive analytics framework. As one of the V’s for characterizing big data, “variety” refers to different forms of big data from various rich big data sources. Our framework collects and integrate data to be used for analysis and prediction from these various rich big data sources.

Depending on applications, data can be collected from different sources. For example, weather data can be collected from government agencies such as (i) Environment Canada and (ii) US National Centers for Environmental Information (NCEI), which is formerly known as National Climatic Data Center (NCDC). As another example, web-search data can be collected from Google Trends5.

Once the data are collected from a wide variety of big data sources, our framework integrates the collected data. The integrated data are then pre-processed prior to predictive analytics. Specifically, the data are cleaned and filtered to consider only those satisfied the user-specified preference (e.g., time period and city of interest). Afterwards, our framework applies feature engineering to indicate important features. For instance, to detect highly correlated variables, the framework uses Pearson’s Correlation—as shown in Eq. (1)—to compute the correlation $\rho_{X,Y}$ between two random variables $X$ and $Y$:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

5 https://trends.google.com/trends/.
where
• $\text{cov}(X,Y)$ is the covariance of $X$ and $Y$; and
• $\sigma_X$ and $\sigma_Y$ are standard deviations of $X$ and $Y$, respectively.

Then, our framework uses a hybrid of three different kernels—namely, linear, polynomial and RBF kernels, as shown in Eqs. (2), (3) and (4), respectively—for the SVM ensemble model in our framework:

$$K(x, x') = \langle x, x' \rangle$$  \hspace{1cm} (2)  \\
$$K(x, x') = (\gamma \langle x, x' \rangle + r)^d$$  \hspace{1cm} (3)  \\
$$K(x, x') = \exp(-\gamma ||x - x'||^2)$$  \hspace{1cm} (4)

The mean absolute error (MAE) measures the improvement brought by each of these three kernels. The kernel with the smallest MAE is selected. Before fitting the model, all variables were scaled in order to avoid the dominance of one variable over the other.

### 3.2 Demonstration on an Application for Disease Analytics

Let us demonstrate our innovative big data predictive analytics framework by an application for disease analytics. Specifically, we apply the framework to predict dengue cases in San Juan, Puerto Rico, from a wide variety of big data sources.

#### 3.2.1 Data Collection

In this paper, we used big data from a wide variety of big data sources such as the following:

• dengue surveillance data, in years 2004 to 2008 from DengAI (a competition for predicting disease spread)\textsuperscript{6}. The data were originated from the US Centers for Disease Control and Prevention (CDC), US Navy’s Naval Medical Research Unit No. 6 (NAMRU-6), and US Military Health System’s Armed Forces Health Surveillance Branch (AFHSB), in collaboration with US universities.

• environmental and climate data, in years 2004 to 2008 also from DengAI, where the data were originated from the US National Oceanic and Atmospheric Administration (NOAA). These data capture daily temperature, precipitation and humidity.

• web-search data, which were collected from the Google Trends platform for terms related to dengue disease between 2004 and 2008 in San Juan, Puerto Rico. The Google search interest (GSI) or trend is measured by the metric “Interest over time”, which is a ranging from 0 to 100, where 0 implies that term is unpopular or not enough data was collected, whereas 100 implies that the term is extremely popular. Figure 1 shows the Google Trends result for the term “dengue” in San Juan, Puerto Rico.

Note that, although DengAI contains the dengue surveillance data as well as environmental & climate data from 1990 to 2008, Google Trends contain web-search data from 2004 onwards.

\textsuperscript{6} https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/.
3.2.2 Data Integration

Once our framework collected data from the aforementioned three key types of data from different sources, it creates a new field “week of year” and then integrates these data based on the existing field “year” and our newly created field “week of year”. The data were also cleaned and filtered to consider only those satisfied the user-specific preference (e.g., time period and city of interest). After such a data pre-processing step, our framework applies feature engineering on the 225 cleaned and filter observations, which correspond to the 225 weeks within years 2004–2008. Table 1 shows some selected variables (i.e., features) and their sources.

| Variable                           | Data source              |
|------------------------------------|--------------------------|
| Total dengue cases                 | CDC, NAMRU-6, AFHSB     |
| Average temperature                | NOAA                     |
| Precipitation amount               | NOAA                     |
| Humidity percent                   | NOAA                     |
| Google search interest (GSI)       | Google Trends            |

3.2.3 Feature Engineering

Recall from Sect. 2 that there is a lag time (about 9–16 weeks) between the weather conditions and the development of the mosquito. This lag comprehends the time that takes for the egg hatching, larvae period and lastly the mosquito.
born. Hence, we created variations of the precipitation, humidity and temperature variables for each lag and then applied Pearson’s Correlation to detect the lags with the highest correlation with the variable “total dengue cases”. Results show (i) Lag 16 for precipitation and (ii) Lag 14 for both temperature and humidity led to the highest correlation with the total number of dengue cases, which brings the possibility to make predictions 16 weeks in advance (based on precipitation) and 14 weeks in advance (based on temperature).

Recall from Sect. 1 that the dengue incubation period ranges from 4 to 10 days, and the first symptoms appear 2–7 days after the mosquito bite. Hence, we varied lag times between 1 to 4 weeks for the GSI variable because infected people would search on the Internet when they feel the first symptoms. Results show that 1-week lag time (for search trends) led to the highest correlation with the total number of dengue cases, which means data generated in Google Trends on previous week could be used as dengue case prediction for the current week. Figure 2 shows a heat map with all these computed correlations.

Fig. 2. Heatmap with correlations among the independent and dependent variables.

### 3.2.4 Predictive Analytics

Afterwards, our framework conducts the predictive analytics with a hybrid of three kernels in the SVM ensemble model: linear, polynomial and RBF kernels. Predictions made by the kernel with the smallest MAE are returned as output to predictive analytics.

### 4 Evaluation and Discussion

To evaluate our innovative big data predictive analytics framework over hybrid big data sources on an application for disease analytics on dengue in San Juan during 2004–2008, we conducted several experiments.
As our framework incorporates additional features or variables (e.g., average temperature, precipitation amount, humidity percent, Google search interest (GSI)), we performed correlation tests to measure the correlations of these features with the total dengue cases. Results in Fig. 2 reveal the following features that are strongly correlated with the number of dengue cases:

- Google search interest (GSI), with correlation coefficient of 0.62;
- temperature, with correlation coefficient of 0.59; and
- precipitation, with correlation coefficient of 0.42.

On the other hand, the feature “relative humidity” is weakly correlated with the number of dengue cases. Figures 3, 4 and 5 graphically show the trends between the number of dengue cases and each of its three strongly correlated variables. In these figures, the $x$-axes show the week of year—i.e., Weeks 1–52 in each of the five years in 2004–2008), and the $y$-axes show the variables (i.e., total GSI, average temperature (in °C), or total precipitation (in mm) over the same week of year in the five years—in logarithmic scale).
As our framework uses a hybrid ensemble of three kernels, we measured the improvement brought by each of these kernels by computing its mean absolute error (MAE) \cite{19} as shown in Eq. (5):

$$MAE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i - \hat{y}_i|$$ (5)

where

- $y_i$ is the true value of the $i$-th sample, and
- $\hat{y}_i$ is the corresponding predicted value.

**Table 2.** MAE comparisons of three SVM kernels and two different set of features

| Set of features          | Kernels          |
|-------------------------|------------------|
|                         | Linear | Polynomial | RBF   |
| \{Temperature, Precipitation, GSI\} | 10.15  | 11.6       | 12.85 |
| \{Temperature, Precipitation\}  | 13.52  | 12.23      | 14.12 |

A MAE value closer to 0 implies fewer errors on the prediction. Here, we evaluated the benefits of incorporating the GSI as a feature in predictive analytics. Results in Table 2 reveal that:

- the presence of the GSI feature improved the linear kernel in the SVM (by 3.37 in terms of MAE)
- the presence of the GSI feature also improved the polynomial and RBF kernels in the SVM.

Thus, the **presence of the GSI feature improved the prediction made by the hybrid SVM ensemble model** (with three kernels). Moreover, we also compared the MAE of predictions made by different kernels. Results in Table 2 reveal that:

- In the presence of GSI features, linear kernel led to the smallest MAE.
- In the absence of GSI features, polynomial kernel led to the smallest MAE.
5 Conclusions

In this paper, we presented an innovative big data predictive analytics framework over big data sources for weather and web-search data. We evaluated and showed that the presence of these data as additional features or variables improved the prediction accuracy by lowering the MAE of the hybrid SVM ensemble model, which consists of three different kernels. Results also show that the linear kernel with the GSI feature led to the most accurate prediction (with the smallest MAE). An application of our framework for disease analytics demonstrates the benefits of our framework to health agencies on the prediction and control of dengue cases. As for ongoing and future work, we plan to apply transfer learning [14] for disease analytics of other diseases and/or at other locations with similar climate characteristics and Internet usage. As a general framework, it can be apply to predictive analytics for other purposes in a wide variety of domains (e.g., prediction of box office [15] or stock market [18]). We also consider to extend our framework as to deal with novel features of big data such as performance, privacy, and flexible paradigms (e.g., [2,5,6,9]).

Acknowledgements. This project is partially supported by NSERC (Canada) and University of Manitoba.

References

1. Audu, A.A., Cuzzocrea, A., Leung, C.K., MacLeod, K.A., Ohin, N.I., Pulgar-Vidal, N.C.: An intelligent predictive analytics system for transportation analytics on open data towards the development of a smart city. In: CISIS 2019. AISC, vol. 993, pp. 224–236 (2019)
2. Chatzimilioudis, G., Cuzzocrea, A., Gunopulos, D., Mamoulis, N.: A novel distributed framework for optimizing query routing trees in wireless sensor networks via optimal operator placement. JCSS 79(3), 349–368 (2013)
3. Chan, E.H., Sahai, V., Conrad, C., Brownstein, J.S.: Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. PLoS Negl. Trop. Dis. 5(5), e1206:1–e1206:7 (2011)
4. Cox, T.S., Hoi, C.S.H., Leung, C.K., Marofke, C.R.: An accurate model for hurricane trajectory prediction. In: IEEE COMPSAC 2018, vol. 2, pp. 534–539 (2018)
5. Cuzzocrea, A., Bertino, E.: Privacy preserving OLAP over distributed XML data: a theoretically-sound secure-multiparty-computation approach. JCSS 77(6), 965–987 (2011)
6. Cuzzocrea, A., Moussa, R., Xu, G.: OLAP*: effectively and efficiently supporting parallel OLAP over big data. In: MEDI 2013, pp. 38–49 (2013)
7. da Silva, M.V.B., Jacobs, A.S., Pfütscher, R.J., Granville, L.Z.: Predicting elephant flows in internet exchange point programmable networks. In: AINA 2019, pp. 485–497 (2019)
8. Dierckens, K.E., Harrison, A.B., Leung, C.K., Pind, A.V.: A data science and engineering solution for fast k-means clustering of big data. In: IEEE TrustCom-BigDataSE-ICESS 2017, pp. 925–932 (2017)
9. Eom, C.S., Lee, C.C., Lee, W., Leung, C.K.: Effective privacy preserving data publishing by vectorization. Inf. Sci. (2019). https://doi.org/10.1016/j.ins.2019.09.035
10. Guo, P., Liu, T., Zhang, Q., Wang, L., et al.: Developing a dengue forecast model using machine learning: a case study in China. PLoS Negl. Trop. Dis. 11(10), e0005973:1–e0005973:22 (2017)
11. Hii, Y.L., Zhu, H., Ng, N., Ng, L.C., Rocklo, J.: Forecast of dengue incidence using temperature and rainfall. PLoS Negl. Trop. Dis. 6(11), e1908:1–e1908:10 (2012)
12. Husin, N.A., Salim, N., Ahmad, A.R.: Modeling of dengue outbreak prediction in Malaysia: a comparison of neural network and nonlinear regression model. In: ITSim 2008, pp. 6–9 (2008)
13. Kesorn, K., Ongrak, P., Chompoosri, J., Phumee, A.: Morbidity rate prediction of dengue hemorrhagic fever (DHF) using the support vector machine and the Aedes aegypti infection rate in similar climates and geographical areas. PLoS ONE 10(5), e0125049:1–e0125049:16 (2015)
14. Leung, C.K., Cuzzocrea, A., Mai, J.J., Deng, D., Jiang, F.: Personalized DeepInf: enhanced social influence prediction with deep learning and transfer learning. In: IEEE Big Data 2019 (2019)
15. Leung, C.K., Eckhardt, L.B., Sainbhi, A.S., Tran, C.T.K., Wen, Q., Lee, W.: A flexible query answering system for movie analytics. In: FQAS 2019. LNCS (LNAI), vol. 11529, pp. 250–261 (2019)
16. Mai, M., Leung, C.K., Choi, J.M.C., Kwan, L.K.R.: Big data analytics of Twitter data and its application for physician assistants: who is talking about your profession in Twitter? In: Data Management and Analysis, pp. 17–32 (2020)
17. Milinovich, G.J., Williams, G.M., Clements, A.C.A., Hu, W.: Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect. Dis. 14(2), 160–168 (2014)
18. Morris, K.J., Egan, S.D., Linsangan, J.L., Leung, C.K., Cuzzocrea, A., Hoï, C.S.H.: Token-based adaptive time-series prediction by ensembling linear and non-linear estimators: a machine learning approach for predictive analytics on big stock data. In: IEEE ICMLA 2018, pp. 1486–1491 (2018)
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., et al.: Scikit-learn: machine learning in Python. JMLR 12, 2825–2830 (2011)
20. Sarumi, O., Leung, C.K.: Scalable data science and machine learning algorithm for gene prediction. In: BigDAS 2019, pp. 118–126 (2019)
21. Yusof, Y., Mustaffa, Z.: Dengue outbreak prediction: a least squares support vector machines approach. IJCTE 3(4), 489–493 (2011)