IMPROVING THE EPIDEMIOLOGY USING THE BIG DATA ANALYSIS WITH SATISTICAL MODELS

Selvarajah Selvendra[1], T.M.J.A Cooray[2]

[12]*Department of Mathematics, University of Moratuwa*

TNTRODUCTION

The popularity of social media with improved performance in analytics will trigger the benefit of using large data sets to other industries as well. The advancement of the IT technologies has resulted in industries moving towards the "big data" analysis rather than the sample techniques. In day to day life, new diseases are evolving with complex issues where some require root cause analysis, therefore preventing diseases should be a very important task to be handled effectively and vigilantly so as to accommodate world changes.

According to the medical research continuity, there is a heavy cost to be borne by individuals to carry out the research, unfortunately most of the funding organization focus on their own interest. In this context some automated systematic approach should be developed to find the risk factors and its changes. Mathematics and IT have long been an important tool for understanding and controlling diseases on a large scale.

The medical arena epidemiology unit is tasked with disease preventing, monitoring and controlling. The epidemiology was first developed to discover and understand the possible causes of contagious diseases like smallpox, typhoid and polio among humans. It has been expanded to include the study of factors associated with non-transmissible diseases like cancer, and of poisoning caused by environmental agents. Epidemiologists determine risk factors associated with diseases and protect people from those diseases. Epidemiological studies can never prove causation; that is, it cannot prove that a specific risk factor actually causes the disease being studied. Epidemiological evidence can only show that this risk factor is associated with a higher incidence of disease in the population exposed to that risk factor.

METHODOLOGY

The methodology illustrates the information clustering and diseases risk factor identification to create a common model to predict individual risk. It predicts the individual risk level based on the possible significant diseases which effect health or economically. Further analyses provide the optimization of the factors to reduce individual risk levels using various BIG data clustering techniques to make the healthcare information a defined structure to analyze. The methodology explains a common way to create a model and its automated approach to predict the risk level of disease, which impacts an individual according to the current information or changing factor patterns.

Step1: Initial risk factors have been selected from past research and other accepted sources which continue analysis by system and adjust based on the historical information and ongoing data changes. It derives the new factors dynamics without performing the additional sampling or guesses. Use the various factor analysis techniques to eliminate the non-significant factor from the model creation. Non-significant factors hold on for future evaluation based on the new records included in the system.

Step 2:Qualitative factors are reclassified further into sub levels to express mathematical models which can be automated and quantitatively measured to control in the future. Residual data sets that are

[1] selvendra@gmail.com 0772245911,0112727121

impacted by disease that are not identified by the model and this data set can be further analysis to find out model variation including new factors.

$$f_j = a_0 + a_i w_1 + \cdots + a_j w_k + e_j, \quad \text{where } e_j - \text{iild}(0,1)$$

k - no of factors which are accepted by the above hypotheses
f_j – j^{th} factor ,
w_i – Sub Factor of fj where i=1,2 ….K

The weightage of the factor are calculated using probability function a_i =P(fi) = n(fi)/n(Dj) and based on the correlation between the factors should be considered to be as separate as possible. Further cluster the factor risk level of the disease to group to minimize the variation within the cluster.

Step 3: Factor progressive of individual's values to predict using the time series mathematical model with historical information of individuals. This time series model iteratively validates to change according to the residuals. This considers changing the value of person and to determine the functions.

$$F_{jt} = c_0 + c_{j1} f_{jt-1} + \cdots + b_{jk-n} f_{jt-n} + e_j, \quad \text{where ej} - \text{iidN}(0,1)$$

F_{jt}- thejth factor influence at time t
C_i – factor coefficient
e_i - Residuals (not related to the m factors)

Until residuals become to iid model adjust automatically and find out more factors from residual data set. Individual Risk Level = Max (f_{jt}, f_{jt-1}……. f_{j1}) where t – determined by Disease Incubation time and other medical delay periods which provide the early identification.

Step 4: Create a Common Model for a Disease: This mainly considers the common way of the disease's and which patterns to determine the functions based on the factor value.

$$Dj = a_{j0} + a_{j1} f_{j1} + \cdots + a_{jk} f_{jk} + e_j, \text{ where ej} - \text{iidN}(0,1)$$

Dj - Disease impact risk level predicated for single human factors
fji = the ith factor influence to the jth Disease
aji factor jth influence weightage
k = the number of factors
ei - Residuals (not related to the m factors)

Until residuals become to iid model adjust automatically and find out more factors from residual data set. Model has the two parts one is specific to diseases which is common for all and other one specific for an individual. Let's consider the a factor information Ωt

Risk Level for Disease = p*Common Disease Model (Ωt) + (1-p)*Person Specific Model (Ωt)

Ωt - Information at t
P- probability of information accuracy of individual
a_i factor jth influence weightage for Di

P is the probabilities of the information accuracy that are calculated by the system on the individual and system records

Factor optimization

Each disease has some set of factors, some are common for some diseases. Humans are more concerned about the major diseases which physically or financially impact them, which should optimistically reduce. System calculates the most important influence factor that gives optimum risk level reductions.

$$fj = \sum \alpha_i * D_i$$

D_i - Disease impact risk level predicated for single human factors
f_j = the jth factor influence
a_i factor j_{th} influence weightage for D_i

RESULT AND DISCUSSION

Death should be a common fear for every human; it's used by some of the companies as an advantage to perform as business.

When consider lung cancer to discuss concept of analysis using statistical models for explanation. The first step system that should be done is to filter data from BIG data and who is effected by lung cancer in the past which is the total cancer population in the system which undergoes various cluster and factor analysis to create the models.

The world age-standardized rate (ASR) for lung cancer is 22.9 per 100,000 populations. The estimated lung cancer population is around 1.63 million among the estimated to world population of 7.122 billion. (Source: Wikipedia)

WHO has classified lung cancer factors, let's assume that these below factors are found by the system using above mentioned analysis.

F1 -Tobacco use
F2 -Being overweight or obese
F3- Unhealthy diet with low fruit and vegetable intake
F4- Lack of physical activity
F5- Alcohol use
F6- Sexually transmitted HPV-infection
F7- Urban air pollution
F8- Indoor smoke from household use of solid fuels.

Behavioral factors (F1, F2, F3, F4, F5, and F6) can be controlled by the individual according to information and its change. However the above the factors are not detailed enough to differentiate the people and risk levels since there are differences person to person. Therefore there should be some more information required to carry out the research. F3 – Healthy diet is not specific to analysis, it should be further sub divided to identify exact factors which impact lung cancer. Factor F2 is a quantitative value which is able to analyze common and individual progressive factors to provide the factor model

Environmental factors (F7, F8) are not controlled by individuals, however they can take some decisions according to the information available and also further sub divide to analyze how government/ communities react to those factors.

Identified factors to time series analysis will provide time variant information to model to respective factor which help to predicate in the future the factor's that changed based on the parameters in the BIG data.

This approach takes some time to accurately model and to record the necessary historic information which can be analyzed and to predict the model. However this is the initial steps for a large healthcare system which analysis should evolve with the changes.

CONCLUSION

The study is considerate of relationships between the epistemology, mathematical and "BIG Data" analysis which improves the analysis of risk factors and help to optimum elimination value. Further explains the conceptual analysis of diseases and data gathering. Since the system considers a large data set which improves the accuracy of system and its self-providing capabilities to correct residual which helps to correct the model automatically. Further analysis required to provide the final conclusion.

REFERENCE

Ambiga Dhiraj, Michele Chambers, Michael Minelli, Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses

Bonita R, Beaglehole R, Kajallstrom T, Basic epidemiology 2nd EditionWHO Press, World Health Organization.

Cooray T.M.J.A, Applied Time Series Analysis and Forecasting

Gintautas DZEMYDA, Leonidas SAKALAUSKAS Large-Scale Data Analysis Using Heuristic Methods

Mike Barlow, Real-Time Big Data Analytics: Emerging Architecture

Mike Ferguson, Architecting A Big Data Platform for Analytics

Robin Bloor, Ph D, BIG DATA ANALYTICS - THIS TIME IT'S PERSONAL

RUEY S.TSAY, Analysis of the Financial Time Series, A John Wiley & Sons, Inc.