SSARSC

Scientific Society of Advanced Research and Social Change
SSARSC International Journal of ICT
Volume 2 Issue 1, January-June 2020, ISSN 2581 - 5873

CANCER PREDICTION USING SPATIAL DATA MINING

Palak Srivastava[1],Deepshikha[2],Deeksha Gangwar[3], S.N Rajan[4],Monalisa Panigrahi[5]

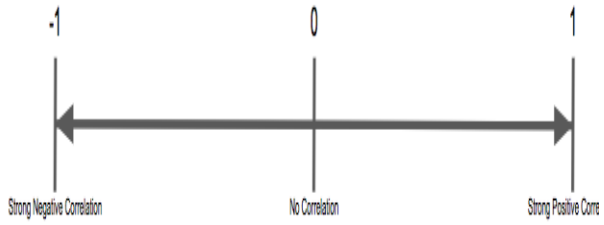
Department of Information Technology,
IMS Engineering College, Ghaziabad, India
(deepshika.singh2520@gmail.com)
(palaksrivastava67@gmail.com)

***ABSTRACT*— Identifying cancer cases by proposing a novel co-relation algorithm. In this context, we specifically attemptto understand whether there is a relationship between the prevalence of cancer and certain factors like Pollution Level(PMindex),Obesity,Alcohol intake in a particular Location.Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. Among several types of correlation coefficients Pearson's correlation(also called Pearson's R) which is a correlation coefficientcommonly used in linear regression have been used for the implementation process.**

Keywords-Data Mining,CorelationCofficient,PM Index

I.INTRODUCTION

The Correlation Coefficient is a widely used method of determining the strength of the relationship between two numbers or two sets of numbers. This coefficient is calculated as a number between -1 and 1. 1 being the strongest possible positive correlation and -1 being the strongest possible negative correlation.



A positive correlation means that as one number increases the second number will also increase.A negative correlation means that as one number increases the second number decreases.Whether or not the outcome of the second number is CAUSED by the first is not being determined here, just that the outcomes of the two numbers happen in concert with each other.If the formula returns 0 then there is absolutely NO correlation between the two sets of numbers.

II.STUDY FACTORS

1.Pollution Level(PM 10):Chronic exposure to particles contributes to the risk of developing cardiovascular and respiratory diseases, as well as of lung cancer. Air quality measurements are typically reported in terms of daily or annual mean concentrations of PM_{10} particles per cubic meter of air volume (m^3). Currently, the WHO identifies safe levels of PM10 - particulate matter measuring under ten micrometres - as under 20 micrograms per cubic metre. This is much lower than the EU's safe particulate matter level, which stands at 40 micrograms per cubic metre.

2.Obesity:Evidence that obesity is associated with cancer incidence and mortality is compelling. By contrast, the role of obesity in cancer survival is less well understood. There is inconsistent support for the role of obesity in breast cancer survival, and evidence for other tumor sites is scant. The variability in findings may be due in part to comorbidities associated with obesity itself rather than with cancer, but it is also possible that obesity creates a physiological setting that meaningfully alters cancer treatment efficacy. In addition, the effects of obesity at diagnosis may be distinct from the effects of weight change after diagnosis.

Obesity and related comorbid conditions may also increase risk for common adverse treatment effects, including breast cancer–related lymphedema, fatigue, poor health–related quality of life, and worse functional health. Racial and ethnic groups with worse cancer survival outcomes are also the groups for whom obesity and related comorbidities are more prevalent, but findings from the few studies that have addressed these complexities are inconsistent.

3.Alcohol Intake:Alcohol consumption has been linked to an increased risk for various types of cancer. A combined analysis of more than 200 studies assessing the link between alcohol and various types of cancer (i.e., a meta-analysis) sought to investigate this association in moredetail. This meta-analysis found that alcohol most strongly increased the risks for cancers of the oral cavity, pharynx, esophagus, and larynx. Statistically significant increases in risk also existed forcancers of thestomach, colon, rectum, liver, female breast, and ovaries. Several mechanisms have been postulated through which alcohol may contribute to an increased risk of cancer. Concurrent tobacco use, which is common among drinkers, enhances alcohol's effects on the risk for cancers of the upper digestive and respiratory tract. The analysis did not identify a threshold level of alcohol consumption below which no increased risk for cancer was evident.

III.METHODOLOGY

Pearson Correlation Coefficient Formula-

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

- N = number of pairs of scores
- Σxy = sum of the products of paired scores
- Σx = sum of x scores
- Σy = sum of y scores
- Σx^2 = sum of squared x scores
- Σy^2 = sum of squared y scores

The strength of the relationship varies in degree based on the value of the correlation coefficient. For example, a value of 0.2 shows there is a positive relationship between the two variables, but it is weak and likely insignificant. Experts do not consider correlations significant until the value surpasses at least 0.8. However, a correlation coefficient with an absolute value of 0.9 or greater would represent a very strong relationship.

Implementaion is performed via PostGIS.

PostGIS is a spatial database. Oracle Spatial and SQL Server (2008 and later) are also spatial databases.PostGIS turns the PostgreSQL Database Management System into a spatial database by adding support for the three features: spatial types, indexes, and functions. Because it is built on PostgreSQL, PostGIS automatically inherits important “enterprise” features as well as open standards for implementation.

PostgreSQL is a powerful, object-relational database management system (ORDBMS). It is released under a BSD-style license and is thus free and open sourcesoftware. As with many other open source programs, PostgreSQL is not controlled by anysingle company,but has a global community of developers and companies to develop it.

Enabling PostGIS:

PostGIS is an optional extension that must be enabled in each database you want to use it in before you can use it. Installing the software is just the first step. DO NOT INSTALL it in the database called postgres

Connect to your database with plsqlorPgAdmin. Run the following SQL. You need only install the features you want:

Step 1:CREATE EXTENSION postgis;

Step 2: CREATE EXTENSION postgis_topology;

Step 3: CREATE EXTENSION postgis_sfcgal;

Step 4: CREATE EXTENSION fuzzystrmatch

Step 5: CREATE EXTENSION address_standardizer;

Step 6: CREATE EXTENSION
address_standardizer_data_us;

Step 7: CREATE EXTENSION
postgis_tiger_geocoder;

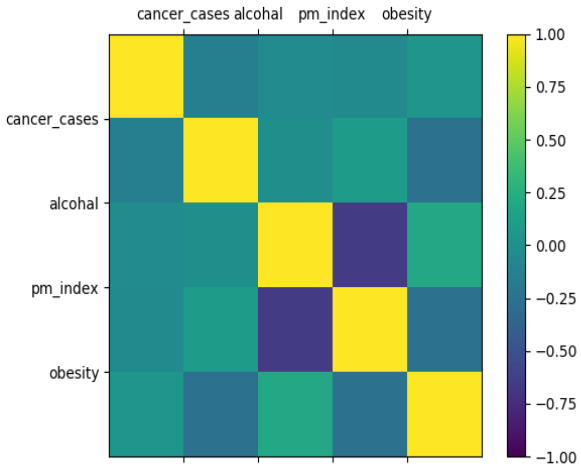
IV.RELATED WORK

1.**On discovering co-location patterns in datasets: a case study of pollutants and child cancers.**

Identifying relationships between cancer cases and pollutant emissions by proposing a novel co-relation mining algorithm. In this context, we specifically attempt to understand whether there is a relationship between the location of a child diagnosed with cancer with any chemical combinations emitted from various facilities in that particular location. Most of the previous works in this domain are based on transaction-free apriori-like algorithms which are dependent on user-defined thresholds,and are designed for boolean data points. Due to the absence of a clear notion of transactions, it is nontrivial to use association rule mining techniques to tackle the co-location mining problem. Our proposed approach is focused on a grid based transactionization of the geographic space, and is designed to mine datasets with extended spatial objects. It is also capable of incorporating uncertainty of the existence of features to model real world.

V.RESULTS:

With the help of Correlation Coefficient we predict the prevalence of cancer in an individual based on several factors like Pollution,Alcohol Consumption and Obesity.



On the given scale of -1.00 to 1.00 we have used different colors to represent the correlation between cancer prevalence and other factors of our study(Pollution level,Obesity and alcohol consumption).

For example- A correlation between cancer cases and pm_index has its value in the range 0.00-0.25.

VI.CONCLUSIONS

One function that can be useful in determining how two numbers relate to each other is the correlation function.

The Correlation Coefficient is a widely used method of determining the strength of the relationship between two numbers or two sets of numbers. This coefficient is calculated as a number between -1 and 1. 1 being the strongest possible positive correlation and -1 being the strongest possible negative correlation.With the help ofCorrelation Coefficient we predict the prevalence of cancer in an individual

based on several factors like **Pollution,Alcohol Consumption and Obesity**.

Thus we conclude that:

1.There is positive correlation between cancer cases and pm index value,between cancer cases and alcohol consumption ,between cancer cases and obesity.

2.There is no correlation between alcohol and pm index,pm index and obesity,obesity and alcohol etc.

VII.ACKNOWLEDGEMENT

This research was supported by [Associate Professor and Head Of Department Dr.S.NRajan(Information Technology),Ghaziabad]. We thank our mentor [Asst Professor Mrs MonalisaPanigrahi,IT Department Ghaziabad] who provided insight and expertise that greatly assisted the research. We would also like to show our gratitude to them for sharing their pearls of wisdom with us during the course of this research. We are also immensely grateful for their comments on an earlier version of the manuscript, although any errors are our own and should not tarnish the reputations of these esteemed persons.

VIII.REFERENCES

1.Jundong Li1 · Aibek Adilmagambetov2 · MohomedShazanMohomed Jabbar2 · Osmar R. Za¨ıane2 · Alvaro Osornio-Vargas3 · Osnat Wine3 Received: 16 February 2014 / Revised: 27 October 2015 / Accepted: 28 March 2016 © Springer Science+Business Media New York 2016
[On discovering co-location patterns in datasets: a case study of pollutants and child cancers]

2.ShashiShekar, JS Yoo, J. Smith, J.P. Kumquat, ―A partial join approach for mining co-location patterns‖, proceedings of 12th Annual ACM Workshop,
Washington DC, USA, pp.241-249, 2004.

3.Shashi Shekar, Y. Huang, Lecture Notes on ―Discovering Collocation Patterns‖, Computer Science, Springer, 2001.

4.Sanjay Chawla, Florian Verhein, ―Mining SpatioTemporal Association Rules, Sources, Sinks, Stationary
Regions and Thoroughfares in Object Mobility Databases‖, Technical Report No.574, Oct 2005.

5.https://www.postgresql.org/docs/

6.https://chartio.com/learn/postgresql/using-the-postgres-corr-function/

7.https://www.postgresql.org/developer/

8.https://gco.iarc.fr/today/online-analysis-map?v=2018&mode=population&mode_population=continents&population=900&populations=900&key=asr&sex=0&cancer=39&type=0&statistic=5&prevalence=0&population_group=0&ages_group%5B%5D=0&ages_group%5B%5D=17&nb_items=5&group_cancer=1&include_nmsc=1&include_nmsc_other=1&projection=natural-earth&color_palette=default&map_scale=quantile&map_nb_colors=5&continent=0&rotate=%255B10%252C0%255D

9.https://www.who.int/airpollution/data/cities/en/

10.https://opendata.socrata.com/Government/Alcohol-Consumption-Per-Country/hj43-2bpj

11.https://data.world/health/obesity-by-state-2014

12.https://scholar.google.co.in/scholar?q=Related+works+on+correlation+and+cancer&hl=en&as_sdt=0&as_vis=1&oi=scholart