County-level Geographic Distributions of Diabetes in Relation to Multiple Factors in the United States

Gulzhahan Oraz and Xiao Luo

Abstract—The increasing prevalence of diagnosed diabetes has drawn attention of researchers in recent years. In this study, a feature selection method based on linear regression has been used to identify the most relevant factors that are associated with diabetes prevalence from the national county health ranking data sets. Then, Expectation-Maximization clustering algorithm has been used to identify the geo-clusters of counties based on the factors and their relations to the diabetes prevalence for years from 2014 to 2017. The results have identified the unique county-level geographic disparities and trends in diabetes and the related factors over the past four years.

I. INTRODUCTION

Research studies and statistics have shown that diabetes prevalence rate has been increasing significantly in the world over the last decade, and will continue to increase rapidly over years, reaching 300 million cases in 2025 [1]. Research studies have shown that disparities in the prevalence of geographic regions associated with different factors, such as socioeconomic factors [2], ethnicity [3] and etc. In 2011, Barker et al. [4] published a paper that identified the diabetes belt on the the U.S. map using the Behavioral Risk Factor Surveillance System (BRFSS) data of 2007 and 2008. The diabetes belt was located in the Southeast region. This research analyzed the diabetes prevalence in association with individual factor, such as education, gender, life style, ethnicity and etc. Mokdad et al. [3] analyzed changes in diabetes prevalence from 1990 to 1998 by selected characteristics and by State. The characteristics included age-group, sex, education level, etc. They found that of the 43 states compared, 35 showed an increase in the diabetes prevalence. Michimi and Wimberly [5] used the BRFSS data from 2000 to 2006 to study the geographic patterns of obesity and associated risk factors including physical activity, and fruit and vegetable consumption in the U.S. Although the diabetes prevalence rate was not directly involved in their study, the obesity was detected as one of the major risk factors for type 2 diabetes prevalence. Their results showed that higher prevalence of obesity in the counties of the South, whereas lower prevalence appeared in the West and the Northeast regions.

Most of the existing research used statistical analysis, such as regression analysis or geographically weighted regression (GWR) to determine the correlations and spatial patterns. It investigate the geospatial correlation between diabetes and one factor at a time. To the best of the authors’ knowledge, none of the previous research investigated the geographical variation of diabetes prevalence in relation to multiple factors concurrently by using machine learning algorithm. In this research, we use the Expectation-Maximization (EM) algorithm to identify the county-level geo-clusters in the U.S. based on multiple factors. The factors are selected through applying the wrapper feature selection method based on linear regression with regards to the diabetes values. The objective is to identify multiple factors that contribute the diabetes prevalence of specific regions, so that public health researchers and practitioners can gain a better understanding of diabetes prevalence corresponding to the local population health. These results can also provide reasoning for tailoring policies for intervention. The experimental data sets included in this research are national county health rankings data of years 2014 to 2017 [6]. The results have identified the geographical disparities in diabetes in relation to multiple factors concurrently for each individual year, and also demonstrated notable changes over consecutive years.

The paper is organized as follows. The methodologies are detailed in Section II. Section III presents the experimental results and analysis. Finally, conclusions are drawn and the future work is implied in Section IV.

II. METHODOLOGY

The main objective of this research is to explore whether different factors contribute to the unique geographic patterns of diabetes prevalence in the United States. To this end, first, all factors that have possible correlation with diabetes prevalence were extracted from the data set [6]. The wrapper method based on linear regression was used to select the most relevant factors (features) from the extracted factors. Then, the Expectation-Maximization (EM) algorithm was applied to identify the geo-clusters based on the diabetes prevalence and the selected factors.

A. Wrapper Method for Feature Selection

In data mining, feature selection deploys a learning algorithm to select a subset of features that are relevant to the problem and ignore the rest. One category of the feature selection is the wrapper method. In the wrapper method, the feature subset selection algorithm exists as a wrapper around the induction algorithm. The induction algorithm itself is part of the function evaluating feature subsets. In this research, the induction algorithm is a linear regression algorithm. The selection of the features is to choose \( k \) factors to predict diabetes prevalence via linear regression, which means to minimize the mean squared prediction error \((E)\) between
diabetes prevalence ($Y$) and other factors ($X_i$), as Equation 1 shows.

$$E = (Z - \sum_{i \in k} w_i X_i)^2$$  \hspace{1cm} (1)

B. EM clustering algorithm

The Expectation-Maximization (EM) clustering algorithm [9] has been widely used in the health informatics domains for pattern recognition, especially when the number of clusters is unknown. There are two major steps for the EM algorithm: E-step and M-step. The E-step computes the posterior distribution for all data points \(\{x_1, x_2, \ldots, x_M\}\) by using the estimated hidden variables \(\{C_1, C_2, \ldots, C_N\}\) and the parameter \(\theta\) for the hidden variables. The posterior computing for a given data point \(x_i\) is given in Eq. 2

$$P(C_j|x_i, \theta) = \frac{P(x_i|C_j, \theta)P(C_j, \theta)}{\sum_{n=1}^{N} P(x_i|C_n, \theta)P(C_n, \theta)}$$  \hspace{1cm} (2)

The M-step uses the posterior distribution that is calculated at in the E-step for the hidden variables, and optimize the parameter \(\theta\) using the expected values of the hidden variables, as Eq. 3.

$$\theta' = \text{arg max} P(C|x, \theta)$$  \hspace{1cm} (3)

The E-step and M-step iterate until there are no changes between the parameter \(\theta\) and optimized \(\theta'\) for two iterations or the number of iterations reaches the pre-defined maximum number of iterations.

In this research, the wrapper method based on linear regression and EM algorithm implemented in Waikato Environment for Knowledge Analysis (WEKA) [12] were employed.

III. EXPERIMENTAL SETUPS, RESULTS AND DISCUSSIONS

The data sets for this research paper were obtained from the County Health Rankings Project [6]. Based on the literature review [7] [8] [13], the factors: poor or fair health (PFH), adult smoking, adult obesity, food environment index (FEI), physical inactivity, excessive drinking, diabetic screening, some college (education), unemployment and uninsured adults were related to diabetes to some extent. Then, the wrapper method is employed as feature selection to select the most relevant ones to the diabetes value. Indeed, feature selection is applied to all years’ data combined after data normalization. Other than adult smoking and some college (education), the rest of the features were selected as the most relevant ones.

The EM algorithm has been applied to the data to observe the geo-cluster patterns of each individual year and changes over the years. The clustering results show that year 2014 has 7 clusters, year 2015 has 10 clusters, years 2016 and 2017 have 25 clusters respectively. The total number of clusters significantly increases from year 2015 to 2016. The possible reason is that the relationships between the factors and diabetes have changed for some regions because of the changing of the local policies or economic situations. Hence, the size of clusters become smaller, and the clusters reflect local characteristics of the population health more precisely.

Figure 1 shows three notable geo-clusters among six for year 2014 on the map. Table I includes the mean and standard deviation values for these clusters. The cluster 5, which is colored as red on the map, has the highest diabetes value along with all other closely related factors which include: the highest unemployment, the highest food environment index, the highest adult smoking, the highest poor or fair health value, the lowest screening value, the highest uninsured value and the lowest food environment index. It is the ‘unhealthy’ cluster that contains counties at the border of Arkansas and Mississippi, South Carolina, Southern region of Alabama, border of Utah and Arizona and some of Texas. The cluster 6 which is colored as dark blue on the map has the highest diabetes screening, the highest excessive drinking value, the highest food environment index value, the lowest uninsured value, the lowest unemployment, but median level of diabetes value. These counties are located at the middle North and Northeast coast. Cluster 4 colored in green contains many counties in Colorado, some in California and some in the New England region. It is the ‘healthy’ cluster that has the lowest diabetes value along with factors that are associated with, such as the lowest physical inactivity value, the lowest adults obesity value and the lowest poor or fair health value.

Figure 2 shows four notable geo-clusters among ten for year 2015 on the map. Table II includes the mean and standard deviation values for these clusters. The cluster 8 is the ‘unhealthy’ cluster. Most of these counties used to be in the cluster 5 of year 2014. That means there is no change in the factors and diabetes values. The counties in cluster 0 colored as darker blue on the map representing the median level of diabetes, the lowest unemployment, the lowest uninsured value, the highest food environment index and the highest diabetes screening are located at Northeast coast and central North regions. These counties are in cluster 6 for year 2014. The number of counties in this cluster has decreased. The counties in cluster 9 are colored as red. Many of them are in cluster 5 of year 2014. However, since the unemployment value and the physical inactivity value have decreased, this cluster is only characterized by the highest uninsured, the lowest diabetes screening but median diabetes value. However, it is interesting that the uninsured rate has not decreased along with the decreasing of the unemployment rate. Cluster 7 colored as green is the ‘healthy’ cluster. Many of the counties have been in this cluster since 2014. However, a small number of counties no longer belonged to this ‘healthy’ cluster in 2015, such as some Western counties in Colorado.

Starting from 2016, more diversity of clusters can be observed. Figure 3 and tables III show the six notable geo-clusters out of 25 for year 2016. The cluster 22 is the ‘unhealthy’ cluster. The distribution is similar to the cluster 8 of year 2015 but a smaller number of counties. The cluster 10 which is colored as light purple on the map has the highest diabetes screening value, the highest excessive
drinking value, but median level of diabetes value. These counties are located at the middle North. The counties (in cluster 23) that have median level of diabetes value and the highest food environment index value are located at Northeast coast and are colored as darker blue on the map. The counties (in cluster 1) that are colored as red have the highest uninsured value but the second highest diabetes value. With the exception of one county is in North Carolina, all the remaining are in Texas. Cluster 12 is the ‘healthy’ cluster. Although the characteristics of this ‘healthy’ cluster remains the same, the number of counties decreased significantly. The counties in cluster 16 which used to be in cluster 9 in 2015 are colored as brown. However, since the uninsured value has improved, they are separated from the original cluster, yet the diabetes screening still is the lowest. This implies that despite insurance rate increases the diabetes screening rate is still low. Specific further research and intervention are needed for that area.
evaluation methods, and investigating the associations of the includes evaluating the clustering results through internal prepare diabetes prevention plans accordingly. Future work be used by local policy makers and health care providers to increase the diabetes screening rate. The overall findings can for some regions, decreasing the uninsured rate might not generally increasing the diabetes screening decreases the are determined by diabetes and multiple related factors. This might indicate increasing the screening can reduce the diabetes prevalence for those regions. The counties in cluster 6 are similar to cluster 23 of year 2016. The number of counties increased, but there was no significant change in the # of Counties 780 155 504 617 279 129 PFH mean 0.29 0.30 0.19 0.12 0.16 0.24 std. dev. 0.05 0.03 0.01 0.01 0.02 0.04 obesity mean 0.283 0.405 0.302 0.275 0.180 0.359 std. dev. 0.02 0.04 0.02 0.02 0.03 0.04 FEL mean 6.68 3.35 8.32 8.40 7.66 4.34 std. dev. 1.71 1.18 0.48 0.54 0.90 2.06 unemployment mean 0.08 0.10 0.038 0.043 0.035 0.082 std. dev. 0.04 0.02 0.008 0.007 0.005 0.03 inactivity mean 0.240 0.340 0.212 0.210 0.132 0.264 std. dev. 0.014 0.023 0.024 0.015 0.024 0.028 drinking mean 0.1465 0.105 0.222 0.186 0.194 0.174 std. dev. 0.0166 0.0129 0.0215 0.013 0.036 0.030 uninsured mean 0.310 0.221 0.086 0.113 0.146 0.287 std. dev. 0.072 0.035 0.015 0.033 0.042 0.056 screening mean 0.832 0.815 0.896 0.878 0.841 0.559 std. dev. 0.046 0.051 0.030 0.021 0.052 0.1927 diabetes mean 0.097 0.170 0.085 0.096 0.055 0.122 std. dev. 0.011 0.020 0.009 0.008 0.010 0.032 factors for the specific clusters.

REFERENCES

6. County Health Rankings & Roadmaps program, http://www.countyhealthrankings.org/about-project
7. J. Aaron Hipp and N. Chalise, Spatial Analysis and Correlates of County-Level Diabetes Prevalence, 2009 - 2010, Preventing chronic disease, 2015, 12
11. R Sridharan, Gaussian mixture models and the EM algorithm,

Figure 4 and tables IV show the seven particular geo-clusters out of 25 for year 2017. The cluster 3 colored as light blue corresponds to the cluster 22 of year 2016, and they have very similar distributions. However, the number of counties within that cluster increases from 2016 to 2017. The cluster 5 colored as light purple corresponds to the cluster 2 of 2016. It was observed that the diabetes value for this cluster has decreased along with the increasing of the screening value. This might indicate increasing the screening can reduce the diabetes prevalence for those regions. The counties in cluster 6 are similar to cluster 23 of year 2016. The number of counties increased, but there was no significant change in the distribution. The counties (in cluster 1) colored as red had the highest uninsured value but the median diabetes value. Similar to cluster 1 of year 2016, the number of the counties increases, especially more in Texas, Utah and California. Cluster 12 colored in green is the ‘healthy’ cluster. There is a slight increase in the number of counties in this cluster in 2017, especially in Colorado. The counties (in cluster 20) colored as brown correspond to the ones in cluster 16 of year 2016, however, the screening rate has increased slightly.

IV. CONCLUSIONS AND FUTURE WORK

Through this study, we have identified geo-clusters that are determined by diabetes and multiple related factors. Generally, increasing the diabetes screening decreases the diabetes value of a cluster. However, the results show that for some regions, decreasing the uninsured rate might not increase the diabetes screening rate. The overall findings can be used by local policy makers and health care providers to prepare diabetes prevention plans accordingly. Future work includes evaluating the clustering results through internal evaluation methods, and investigating the associations of the