Classification Analysis of Children’s Blood Lead Level in Syracuse, NY

Final Project of Spring 2016 Data Mining

Qianqian Cao

Introduction

Lead exposure has adverse impacts on children’s nervous systems, resulting low IQ and learning disabilities (Lanphear 2000). No safe blood lead level (BLL) in children has been identified. Even low-level lead in blood has been shown to influence IQ, ability to pay attention, hearing, and learning problems and slows the child’s growth (Chiodo et al. 2004). Therefore, protecting children from exposure to lead is important to lifelong good health.

Children under six years of age have the highest exposure to lead due to behavior and other factors, such as greater hand dust contamination, frequent hand-to-mouth transfer and higher absorption rates than adults. The major environmental lead sources are from lead-based paint, lead in indoor dust and bare soil, as well as lead in water (Markowitz 2000).

Socio-economic factor are considered as another indicators of vulnerability to children’s lead poisoning. Households of higher socio-economic status usually have access to more solutions to reduce children’s vulnerable lead poisoning, while poor families may depend on government intervention and assistance for the lead exposure amelioration.

Objective

In 2012 the Center for Disease Control and Prevention (CDC) reduced the BLL threshold of concern to 5 micrograms per deciliter (μg/dl) from 10μg/dl in order to fully address the extent of the childhood lead poisoning in the US (CDC 2013). At the national level, 2.5% of children under the age of six have BLLs greater than 5μg/dl (i.e., the elevated BLLs) (CDC 2013). However, in some inner cities, the incidence of lead poisoning remains high. For example, in the city of Syracuse, NY, 4903 children under the age of 6 years were tested at least once in 2011, and 168 (3.43%) of these children had BLLs ≥ 10 μg/dl and 900 (18.36%) children ≥ 5 μg/dl.

Modeling the lead concern level based on environmental risk and socio-economic factors would be helpful to target the lead hazard areas for intervention programs, mitigate lead contaminations, and prevent children’s lead poisoning. Therefore, the purpose of this study is to explore the relationship between of children’s new blood lead concern level and related risk factors in terms of data mining algorithms.

Software R is used to predict the author via both classification and clustering algorithms. Classification algorithms of Decision Tree, Naïve Bayes, and Support Vector Machines (SVMs) are chosen to classify.

Data Description

In this study, the surveillance data of the children's blood lead levels (BLL) were provided by the Onondaga County Health Department (OCHD). The dataset included the children's lead screening tests from 2007 to 2011 in the city of Syracuse, New York, USA. To avoid the bias caused by follow-ups, only the first test of each child was chosen in the surveillance dataset. Then the blood lead test database is combined with the census block map layer, and found that a total of 1393 census blocks have children lead screening records.

Based on the new elevated blood lead level (≥ 5 μg/dl), each census block was recorded as no child with the BLL ≥ 5 μg/dl (coded as 0), or if there were children with the BLL ≥ 5 μg/dl (coded as 1), which produced a binary response variable, indicating the incident rate of children's lead poisoning.

On basis of previous study and data availability, the building year, town taxable values, soil lead concentration, population count, house number, and area within each census block are numerous values selected as environmental and socio-economic attributes for classification analysis. Also, the coordinates of block centroid are involved as spatial factor.

Values of location coordinates of geometric center of block area, building year, town taxable value, population, and house counts are provided by the reference data of US Census Bureau. The building year and town taxable values of the houses in a census block were then averaged in ArcGIS 10 to get the value at the census block level. The soil lead concentration was based on the work of Johnson who collected and analyzed 3000 soil samples across the city of Syracuse in the summer of 2003 and 2004 (Johnson and Bretsch.J.K. 2002).

Data Preparation

1. Removing inconsistent record

Because data resources are provided by different organizations, Onondaga County Health Department and US Census Bureau, it is important to check and remove the inconsistent examples. For example, if the population or house count in one block is zero, but the record of children's blood lead level from OCHD exists, then the example not consistent. A total of 46 inconsistent examples are removed, so final examples are 1347. Also, Id variable is removed in order to prevent over-fitting problem.

2. Preparing train and test dataset

Data is randomly divided into train and test datasets with ratio of 4 to 1. Training data includes 1077 examples, while 270 records are distributed to test part.

Results

1. Information Gain

Table 1 demonstrated the information gain contributions of each variable. Soil lead concentration, building year, and tax value are the top three factors of incident rate of children blood lead level. One interesting phenomenon is the latitude, which is the fourth important factor, but Longitude is not significant at all. It indicates that there is difference between east and west in Syracuse, matching the special phenomenon of economic difference along No. 81 highways.

Table 1. Information gain values of variables for incident rate of children's BLL.

House Number	Population	Latitude	Longitude	Tax Value	Building Year	Block Area	Soil Lead
0.07926	0.09030	0.09375	0.00000	0.1071	0.1204	0.01367	0.1361

2. Decision Tree

Pruned decision tree is listed below. The structure implies that the soil lead concentration is the root node, while house number, Latitude, and tax value and other variables are internal nodes. According to the decision tree, the soil lead and Latitude are positively related to the happen of children blood lead level, while building year and tax value are negative.

Comment [T1]: How do you tell Latitude is positively relate to the happen of children blood lead level? When longtitude > 476940, the class is predicted as 0.

```
DT MODEL
m_dt=J48(COUNT_5~.,data=train,control=Weka_control(C=0.05,M=3))
m_dt
## J48 pruned tree
## ------------------
## Soil_Lead <= 113.396554
## |   House_Number <= 33
## |   |   Latitude <= 410784.4878
## |   |   |   Soil_Lead <= 59.002334: 0 (47.0/4.0)
## |   |   |   Soil_Lead > 59.002334
## |   |   |   |   Tax_Value <= 55323.80952
## |   |   |   |   |   Longitude <= 4769440.938
## |   |   |   |   |   |   Building_Year <= 1946.95122
## |   |   |   |   |   |   |   House_Number <= 5: 0 (3.0)
## |   |   |   |   |   |   |   House_Number > 5: 1 (37.0/9.0)
## |   |   |   |   |   |   Building_Year > 1946.95122: 0 (4.0)
## |   |   |   |   |   Longitude > 4769440.938: 0 (5.0)
## |   |   |   |   Tax_Value > 55323.80952: 0 (93.0/27.0)
## |   |   Latitude > 410784.4878: 0 (34.0)
## |   House_Number > 33
## |   |   Tax_Value <= 62021.42222: 1 (71.0/15.0)
## |   |   Tax_Value > 62021.42222: 0 (50.0/23.0)
## Soil_Lead > 113.396554: 1 (733.0/158.0)
```

3. Naïve Bayes

R package e1071 is applied to build a naïve Bayes model. Laplace smooth is involved to prevent zero probability. Naïve Bayes model without discretization is demonstrated as below. For attributes of house number, population, Latitude, tax value, and soil lead, variation of mean and standard deviation between groups 1 and 0 are more greater than others (e.g. block area), indicating they are important factors.

```
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace, cross = 3,
##     seed = 123)

## Conditional probabilities:
##    House_Number
## Y       [,1]     [,2]
##   0 30.15385 27.42751
##   1 50.19215 58.01107
```

```
##    Population
## Y        [,1]     [,2]
##   0  65.67308  68.8184
##   1 107.99299 100.2984
##
##    Longitude
## Y      [,1]     [,2]
##   0 4766276 2649.949
##   1 4766410 2331.786
##
##    Tax_Value
## Y       [,1]     [,2]
##   0 69386.65 27385.84
##   1 52504.64 17830.41
##
##    Building_Year
## Y       [,1]     [,2]
##   0 1931.938 17.86052
##   1 1918.850 15.36563
##
##    Block_Area
## Y       [,1]     [,2]
##   0 29579.40 46284.55
##   1 30216.73 43550.39
##
##    Soil_Lead
## Y       [,1]      [,2]
##   0 131.1684  89.38427
##   1 209.9942 111.29742
```

Since the kernel density function is Gaussian distribution, it is better to discretize the variables which are not normal distributed. Histograms of variables “Population”, “Block_Area”, “Soil_Lead”, and “House_Number” show non-normal distributions. To match the normal assumption, those four variables are discretized by equal width. Then a new Naïve Bayes model is built with normal distributed variables.

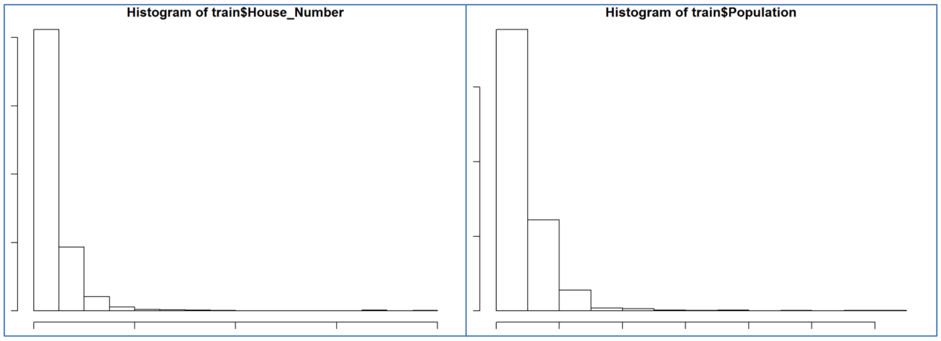


Figure 2. Frequency histograms of House_Number and Population in each block.

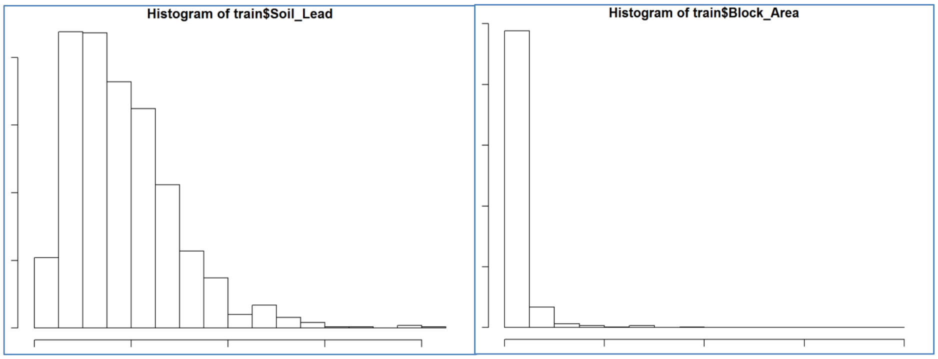
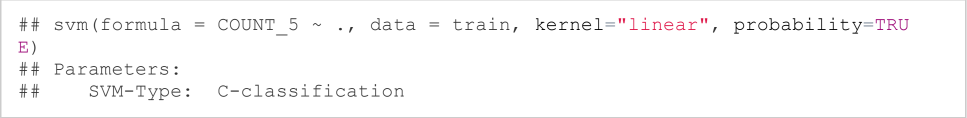


Figure 1. Frequency histograms of Soil_Lead and Block_Area in each block.

4. Support Vector Machines

R package e1071 is also applied to build an SVM model. Linear kernel is chosen because of maximum support vector compared to other kernel functions. Weight vector based on SVM model indicated that variables of population, building year, tax value, and soil lead are more useful.



```
## svm(formula = COUNT_5 ~ ., data = train, kernel="linear", probability=TRU
E)
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  1
##       gamma:  0.125
##
## Number of Support Vectors:  598
##   ( 298 300 )
w <- t(m_svm$coefs) %*% m_svm$SV                 # weight vectors
print(w)
##     Population Building_Year     Tax_Value     Soil_Lead     Longitude
##     -0.84542776   -0.50282049    0.44925095    0.32862775    0.18238290
##   House_Number     Block_Area      Latitude
##     -0.12360716    0.03244900    0.02687689
```

5. Model evaluation

Cross validation is used to evaluate the model performance in train set. Majority baseline is 67.87%. For each model, the number of cross validation folders is fixed to 3. All models are qualified, and SVM has the highest overall accuracy (77.34%). The overall accuracy of each model is summarized in Table 2.

Table 2. Model evaluation performance.

Model	**Overall accuracy %**
Decision Tree	72.42
Naïve Bayes	73.79
Naïve Bayes (discretization)	74.83
SVM	77.34

6. Prediction performance

Four models are applied to predict the test data to identify the block where there is a record of children's BLL ≥ 5 µg/dl. SVM also has the highest prediction performance than other three models (Table 3).

Table 3. Model prediction performances.

Model	**Prediction accuracy %**
Decision Tree	76.67
Naïve Bayes	72.96
Naïve Bayes (discretization)	77.04
SVM	82.96

Discussion

Besides the model evaluation and prediction, the confusion matrix is also important tool to explore deeper information. Since low-level lead in blood has adverse influence to child's growth, it is a lot more cost-effective to prevent children from exposure in risky environment than to treat afterwards. It is expected that the false negative value (actual class is dangerous but predicted as safe place) is as low as possible. Among four models, the Naïve Bayes without discretization has smallest false negative (Table 5), then SVM has the second smallest value (Table 7). Oppositely, the decision tree generates largest false negative (Table 4). So, even though the Naïve Bayes model with discretization performs higher overall accuracy, the default Naïve Bayes model is superior when the cost of false negative is considered.

Table 4. Confusion matrix of Decision Tree prediction

	Predicted Class		
Actual Class	**Decision Tree**	BLL ≥ 5 μg/dl **=Yes**	BLL ≥ 5 μg/dl=**No**
	BLL ≥ 5 μg/dl **=Yes**	165	**48**
	BLL ≥ 5 μg/dl=**No**	15	42

Table 5. Confusion matrix of Naïve Bayes prediction.

	Predicted Class		
Actual Class	**Naïve Bayes**	BLL ≥ 5 μg/dl **=Yes**	BLL ≥ 5 μg/dl=**No**
	BLL ≥ 5 μg/dl **=Yes**	134	**23**
	BLL ≥ 5 μg/dl=**No**	46	67

Table 6. Confusion matrix of discretized Naïve Bayes prediction.

	Predicted Class		
Actual Class	**Naïve Bayes** (discretization**)**	BLL ≥ 5 μg/dl **=Yes**	BLL ≥ 5 μg/dl=**No**
	BLL ≥ 5 μg/dl **=Yes**	159	**41**
	BLL ≥ 5 μg/dl=**No**	21	49

Table 7. Confusion matrix of SVM prediction.

	Predicted Class		
Actual Class	**SVM**	BLL ≥ 5 μg/dl **=Yes**	BLL ≥ 5 μg/dl=**No**
	BLL ≥ 5 μg/dl **=Yes**	168	**34**
	BLL ≥ 5 μg/dl=**No**	12	56

Conclusion

In the inner city, lead poisoning is still a public health with a high proportion of old houses. In this study, four data mining algorithms based on environmental risk and socio-economic factors are used to target the lead hazard areas for intervention programs, mitigations of lead contamination, and prevention of children's lead poisoning. By considering the performances of model evaluation, prediction, and confusion matrix, SVM is the most appropriate model to target lead hazard areas for children's BLL protection.

The findings also suggest that the soil lead concentration, population, tax payment, house number, Latitude and building year are important factors. Features of soil lead concentration and house building year represent the environmental lead hazard exposure. Older houses were painted with high lead level, and they were also associated the plumbing system with lead solder, providing potential lead source for children's lead poisoning. So the block with older house and higher soil lead concentration might not be healthy place for children under 6-year ages. Population, tax payment, house number, and Latitude are factors reflecting socio-economic status. Smaller population, less house number within one block implies higher income and tax payment, and better living environment with less lead exposure. Affluent households are able to remove the lead-based paint in the house at their own expense, or they might also live in newer homes that have less lead exposure. Latitude is also interesting factor due to the special economic pattern from west to east in Syracuse. Latitude location indicates that the blocks on the west side of NO. 81 highways have greater potential of lead hazard than east part. If this model is applied to other inner city that does not has obvious economic trend, the spatial location may not be important issue.

The most important step is to prevent lead exposure before it occurs. Therefore, for the family with younger children, it is very important to evaluate the lead exposure hazard level of living block by considering the issues like house building year, house counts and population. From the aspect of government and related organization, they can use the suggested model to target lead hazard areas for intervention programs and mitigations of lead contamination.

Comment [T2]:
Data Collection: Proficient
Analysis Method: Proficient
Interpretation: Proficient
Conclusion: Proficient

References

Center for Disease Control and Prevention (CDC). 2013. CDC Updates Guidelines for Children's Lead Exposure.

Chiodo LM, Jacobson SW, Jacobson JL. 2004. Neurodevelopmental effects of postnatal lead exposure at very low levels. Neurotoxicol Teratol 26(3):359-371.

Lanphear BP. 2000. Cognitive deficits associated with blood lead concentrations. Public Health Rep 115(6):521 – 529.

Markowitz M. 2000. Lead poisoning. Pediatrics in Review 21(10):327 – 335.

Johnson DL, Bretsch JK. 2002. Soil Lead and Children's Blood Lead Levels in Syracuse, NY, USA. Environ Geochem Health 24: 375-385.