Applying Machine Learning Algorithms to Highway Safety EEPDO

Conference Paper · February 2018

DOI: 10.1109/CSCI.2017.248

5 authors, including:

Di Wu
Jackson State University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Ningning Wang
Jackson State University
10 PUBLICATIONS **27** CITATIONS
SEE PROFILE

Sungbum Hong
Jackson State University
17 PUBLICATIONS **107** CITATIONS
SEE PROFILE

Applying Machine Learning Algorithms to Highway Safety EEPDO

Di Wu[1], Ningning Wang[2], Feng Wang[3], Sungbum Hong[4]
1, 4 CDS&E, 2 Dept. of Mathematics & Statistical Sciences, 3 Dept. of Civil Engineering
Jackson State University
Jackson, MS, U.S.A
1 dwzoon@gmail.com, 2 ningning.wang@jsums.edu, 3 feng.wang@jsums.edu, 4 sungbum.hong@jsums.edu

***Abstract* — Estimating expected equivalent property damage only (EEPDO) is crucial to highway crash hotspot identification (HSID), which is a key component of a highway safety improvement program. During the past 60 years, HSID methodologies advanced steadily from traditional scan based methods to statistical model based methods and have reached the sophistication of encompassing advanced statistical models with many variations and refinements while there still exist a number of theoretical issues unsolved. Consequently, these advanced models are not widely used in the practice of transportation engineering. This paper investigated the performance of an easy to use alternative to estimate the EEPDO -- using machine learning techniques of K nearest neighbor (KNN) algorithm and compared it against the prevalent statistical model -- Negative Binomial (NB). NB assumes that the raw data follow a certain Gamma distribution which is not ubiquitously hold for crash data. Comparatively, being a nonparametric predictor, KNN is expected to produce better estimation on crash data in that it requires no assumption on the raw data. For experiment, a case study was conducted on highway US 49 in Harrison County of Mississippi. The results indicated that KNN outperformed NB.**

***Keywords* -- EEPDO, KNN, NB, machine learning, algorithm.**

I. INTRODUCTION

Due to its extensive societal impact, highway safety has long been emphasized and intensively studied. For the state of Mississippi, the highway safety problem is especially prominent. The traffic safety situation of Mississippi has been severe. For the past 5 years, the traffic fatality per capita of the state stayed at about twice the national average level. Although tremendous executive effort has been made to develop the State Highway Safety Plan, limited attention has been paid to better understanding of the characteristics of the crashes in Mississippi. There is a need to reveal the causation factors for the crashes in Mississippi in terms of but not limited to roadway, traffic, geographic, social, and economic attributes through comparing and identifying the optimum methods of crash data analysis for Mississippi.

Hotspot identification (HSID) is a key component of a highway safety improvement program. For the past 60 years or so, HSID methodologies advanced steadily from traditional scan based methods to statistical model based methods and have reached the sophistication of advanced models with many variations and refinements. Meanwhile, many theoretical issues of the statistical models are not yet solved. Consequently, the advanced models are not widely used in practice. There is a call for the development of an easy to use platform with acceptable accuracy.

A nonparametric approach with machine learning techniques might be an answer. Machine learning emerged in 1990s and has been widely applied in many fields. However, in HSID it was only started in recent years and limited in certain subareas such as clustering. The K nearest neighbors (KNN) algorithm has not been used as an accident prediction model (APM) in HSID although by nature it can be a desire predictor. Being non-parametric, KNN has no assumption on the raw data, which fits the characteristics of crash data. Crash data have the feature of rareness and randomness, which means they do not follow typical and theoretical assumptions. Statistical models, such as Gaussian mixtures, assume the data follow certain probability distribution, which limit their application range. Therefore there is a possibility that the KNN may have a better performance in HSID than the statistical models.

The objective of this paper is to exam the performance of KNN regression in EEPDO of HSID and to develop a platform using the open source software R. This paper is composed of sections of introduction, literature review, methodology, experiment and conclusions.

II. LITERATURE REVIEW

A. HSID

Hot pots are roadway sites with the most potential for crash frequency or severity reduction [1]. A crash is the commonplace word used to describe a failure in the performance of one or more of the driving components, resulting in death, bodily injuries or property damage [2]. Crash is rare and random, fluctuating over time and space.

HSID methods developed with increasing sophistication of techniques are changing in the thinking about road safety [1]. The complexity of the method is expressed by the performance measures it adopts. The performance measures of HSID are related to certain description of crash rate and/or severity, such as crash rate, equivalent property damage only (EPDO), expected average crash frequency with Empirical Bayesian (EB) adjustment etc. The traditional methods adopt observed/scan based measures only, such as crash rate and EPDO, the contemporary methods include expected measures, such as expected equivalent property damage only (EEPDO).

978-1-5386-2652-8/17 $31.00 © 2017 IEEE
DOI 10.1109/CSCI.2017.248

B. Crash Rate Methods

Crashes are rare and random events. Rareness means the circumstances contributed to the crash are scarce compared to a normal condition. Randomness means that crashes occur as a function of a set of events influenced by several factors, which are partially deterministic and partially stochastic. The traditional method ignores the randomness of crash data, assuming the road site with high historical crash rate will continue to have high crash rate in the future. Hot spots are hence identified when the crash rate exceeds the critical threshold, which depends on the average crash rate at similar sites, traffic volume, and a statistical constant that represents a desired level of significance [1]. Similarly, the crash severity method uses the observed crash data only. It includes crash severity by converting all crashes into the numbers of equivalent property damage only crashes on the basis of their severity (fatal, injury, property damage only) or/and the damage caused.

However, when a period with a comparatively high crash frequency is observed, it is statistically probable that this period will be followed by a comparatively low crash frequency [1]. The crash causation factors responsibilities are categorized into human, road, vehicle factors, and combinations of them [2]. The relationship between crash and its causation factors is more complicated than linear. Therefore, when a crash rate method assumes a linear relationship exists between traffic volume and the frequency of crashes, the resulted formula present regressions to the mean bias [1]. This bias cannot be solved unless its fundamental assumption is modified.

C. Expected Crash Rate and Statistical Models

Departure from what is commonly done is difficult to defend [3]. However, methodologies will inevitably change with the improvement of computational capability. As a matter of fact, it was the popularization of personal computers a couple of decades ago that triggered an extraordinary number of researches to apply statistical methods [3].

In order to take account of the fluctuation of crash data over time and space, Empirical Bayesian (EB) Inference was introduced. By specifying the safety of a site as an estimate of its long-term mean instead of short-term count, EB smoothes out the random fluctuation of the crash data [4], [5]. Over the years, a broad variety of statistical APMs were developed for the need of EB implementation.

A Poisson regression model was initially introduced for its capability to address rareness and randomness. Soon, the Poisson model was found not able to solve the over-dispersion of crash data because Poisson distribution assumes the mean equals the variance while for crash data the variance typically exceeds the mean [6]. Therefore, Poisson models were then relaxed to NB models which were used to establish relationships between crashes and highway geometry. The Highway Safety Manual (HSM) 2010 established the HSID procedure based on NB regression models with EB interference.

The extensions of Poisson, NB and EB were widely studied and debated. For example, the multivariate models attempted to explain the correlations between different types of accidents but they were criticized as "empirical better" and the same set of explanatory variables is repeatedly used for each type of response [7]. Zero-inflated Poisson and NB models were developed to simulate the crash rareness with split statues of a crash-free and crash-prone propensity but were questioned for their assumption in that some road segments are always perfectly safe [8]. Markov-switching NB models simulate roadway safety on distinct multi-state representing different conditions [9]. They were found to out-perform single-state models but at the expense of increased complexity.

Many statistical problems are still not yet solved, such as over-dispersion, under-dispersion, time-varying explanatory variables, low sample means and size, crash-type correlation, under-reporting of crashes, omitted-variables bias, and issues related to functional form and fixed parameters [10]. HSM predictive models also have limitations. First, in practice, the predictive models were developed as part of HSM-related researches from the most complete and consistent available datasets. Crash frequencies, even for nominally similar roadway segments or intersections, can vary widely from one jurisdiction to another. Geographic regions differ remarkably in climate, population, driver population, crash reporting threshold, and crash reporting practices [1]. For local jurisdictions, in order to use HSM predictive models, a calibration is needed to address variations between jurisdictions and geographic regions. The accuracy of the calibration method is hard to identify. Therefore, HSM encourages users to replace the default values with locally derived values. Second, in theory, the HSM models incorporate the effects of many, but not all geometric designs and traffic control features of potential interest. It treats the effects of individual geometric design and traffic control features as being independent of one another and ignores potential interactions between them. It is likely that such interactions exist, and ideally, they should be accounted for in the predictive models. At present, such interactions are not fully understood and are difficult to quantify [1].

While the literature shows a steady advancement in crash-frequency analysis with superior statistical fit and/or predictive capabilities [8], with all the limitations of the HSID procedures mentioned, the local agencies are expecting an easy to follow HSID method with acceptable accuracy.

D. Multi-Criteria, Similarity and Machine Learning

Performance measures are a key component of a HSID method. The first step of a HSID would be to select the performance measures. In order to represent many aspects of crash data, HSM provided twelve performance measures to be selected [1]. Because of the formidable complexity, currently, there is a lack of conclusive evidence showing which criterion performs better [10].

KNN is a well-known, easy and successful non-parametric method used for both classification and regression [11], [12]. It is the most basic type of instance learning and widely used in

text classification, image recognition, cancer diagnosis, and economic events prediction. In transportation, KNN was applied in sub-areas of short term traffic volume forecast, vehicle classification, and pavement crack detection.

To the authors' knowledge, a couple of researches applied clustering techniques of machine learning to HSID. For example, a K-means clustering was applied to HSID and resulted in better performance than EB [13], and a comprehensive Fuzzy cluster based HSID platform which also shows better performance than statistical models and easy for implementation [14].

III. METHODOLOGY

In order to exam the performance of KNN and NB model in EEPDO of HSID, in our study, KNN algorithm and the NB model were applied to the same crash data to derive the EEPDO. The evaluation of the performance of the two methods was based on the smallest MSE. The following paragraphs describe the procedures, algorithms.

The crash data, traffic volume and roadway geometry data from different data sources were merged for the study area of interest. The data cleaning, data mapping, and the following statistical analysis and machine learning computation processes were implemented in R and tested on a Windows OS equipped with 3.6G Hz i7 CPU and 16.0G RAM.

A. NB Prediction

The NB distribution is a mixture of Poisson and Gamma distribution. It was developed first by Greenwood and Yule to account for over-dispersion that is commonly observed in discrete or count data [15]. Crash data can be characterized as the product of Bernoulli trails with unequal probability of events (Poisson trails). As the number of trails increases the distribution may follow a Poisson process and the amount of dispersion is governed by the characteristics of this process. Thus, the number of crash (OEPDO) at ith site Yi is assumed to be Poisson distribution with mean and independent over all sites.

$$Y_i|\lambda_i \backsim Poisson(\lambda_i) \quad (1)$$

The Poisson mean λ_i is structured as:

$$\lambda_i = \widehat{\lambda_\iota}\varepsilon_i = f(X;\beta)\cdot\exp(\varepsilon_i), \quad \text{and}$$

$$\lambda_i = \widehat{\lambda_\iota} = f(X;\beta) = \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p) \quad (2)$$

Where X_js are independent variables, represent the total number of independent variables, which are segment length and AADT per lane two independent variables in our case, and ε_i is error term.

For NB model, it is assumed that $\exp(\varepsilon_i)$ is independent and Gamma distributed with a mean equal to 1 and a variance $1/\phi$ for all site i. Under this assumption, it can be shown that Y_i is conditional on $f(\cdot)$ and distributed as a Poisson-Gamma (NB) random variable with mean $f(\cdot)$ and a variance $f(\cdot)(1+f(\cdot)/\phi)$, respectively. Then the probability density function of Y_i described above is given by (3), where y_i is the response variable ith site, λ_i the mean response for observation site i and ϕ is the inverse dispersion parameter of the Poisson-Gamma distribution which is defined as the "inverse dispersion parameter" of Poisson-Gamma distribution [16].

$$P(Y = y_i|\phi,\lambda_i) = \frac{\Gamma(\phi + y_i)}{\Gamma(\phi)y_i!}\left[\frac{\phi}{\phi+\lambda_i}\right]^{\phi}\left[\frac{\lambda_i}{\phi+\lambda_i}\right]^{y_i} \quad (3)$$

After Hauer (1997) examined many accident data sets and the empirical evidence the author obtained supported the gamma assumption for the distribution (18), the use of NB for APM has been a usual practice [17],[18].

A foreign package in R will be used for the NB regression to determine the parameters of a and b, as shown in (4). Using (4), EEPDOs (predicted value of OEPDO) of all sites will be calculated.

$$EEPDO_i = \exp\big(1.7157 + a * segment\ length + b * AADT_{perLane}\big) \quad (4)$$

B. KNN Algorithm

While NB's variations and refinements steadily advancing, major improvement in APM may be achieved by using more appropriate model forms, including nonlinear ones, and possibly using a nonparametric approach [5]. KNN is the simplest nonparametric decision procedure where the function was approximated locally and all computation was deferred until classification [19]. Although simple and intuitive, it has been proved that in a large sample case, KNN rule has a probability of error which is less than twice the Bayes probability of error, and hence is less than twice the probability of error of any other decision rule, nonparametric or otherwise [14]. KNN regression assumes all instances are points in n-dimensional space and the Euclidean distance is measured to determine the "closeness" of the instances. Suppose X is a data set composed of m instances $x_1, x_2, \ldots x_m$ .The Euclidean distance is the ordinary distance between two points in Euclidean space. Let an arbitrary instance x be described by the feature vector as in (5), where a_r (x) denotes the value of the rth attribute of instance x. Then the Euclidean distance between two instances xi and x_j is defined to be d (x_i, x_j), as in (6).

$$x = \{a_1(x), a_2(x), \ldots a_r(x) \ldots a_n(x)\} \quad (5)$$

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n}\left[a_r(x_i) - a_r(x_j)\right]^2} \quad (6)$$

Through the definition, Euclidean distance is dominated by the attributes with wide value ranges, which means that attributes with a wider range will have more influence to the Euclidean distance and it is not necessarily true. Normalization is a commonly used data transformation method to project the raw data onto directions which maximize variances. The simplest method of normalization is to rescaling the range of variable to the range of [0, 1]. The general formula of normalization is given in (7), where x denotes the instances in data set X, x_r' denotes the normalized value of the rth attribute ofx_r.

$$x'_r = \frac{x_r - \min(x)}{\max(x) - \min(x)} \quad (7)$$

Among all the variables of the data set, the target feature is the variable to be predicted-- EEPDO in our case. The whole data set is split to distinct training data and testing data. The training data is the sampled instances from which the algorithm learns, the testing data is the instances whose target feature is to be classified or predicted. KNN regression takes the mean (if the target feature is real-valued) or a majority vote (if the target feature is categorical) of the target feature of the K nearest neighbors identified by the Euclidean distances in the training data to predict the target variable of the test data. The algorithm of KNN is illustrated as follows [20]:

Input:

- D - Training data (subset of X).
- k - Number of neighbors.
- t_i - Input instance to estimate.
- a_r (t_i) - the r^{th} attribute of instance t_i .

Output:

- $\widehat{a_r}(t_i)$ – Estimate of the value of the r_{th} attribute of instance t_i.

KNN algorithm:

- Assign initial values for means μ_1, μ_2,μ_n randomly, where μ_i is the mean of the initial k neighbors.
- Repeat a. Assign each item t_i to the cluster which has the closest mean of the target value of; b. Calculate new mean for each cluster using (8); until convergence criteria is met.

$$\widehat{a_r}(t_i) = \frac{\sum_{j=1}^{k} a_r(x_j)}{k} \quad (8)$$

The choice of the number of neighbor k is critical because the performance of a KNN varies significantly as K is changed when the examples are not uniformly distributed [21]. In theory, when the sample size n approaches infinity, the larger the K, the smaller the error rate, on condition that all the neighbors are close to the testing instance. Since crash is rare and random, this large K and small error rate rule does not apply. Determining the value of K in advance becomes difficult.

Usually in literature, the K parameter in KNN is chosen empirically. The "rule of thumb" is to take the square-root of the training sample size. Mean squared error (MSE) was used to measure the performance to calibrate K value. MSE is a general accepted measure of prediction accuracy. The formula of MSE is shown in (9), where y is the observed value and $\hat{y}$ is the predicted value, n is sample size.

$$MSE = \frac{(y - \hat{y})^2}{n - 1} \quad (9)$$

The KNN predictor is used each time with a different K on the same training data and testing data. Starting from k=1 to k =$\sqrt{n}$, one MSE is calculated for each run. The ultimate k value is choosing such that it has the smallest MSE.

C. KNN Prediction

With K value determined, EEPDOs of all the instances will be estimated through the KNN APM. EEPDO is set as the target attribute. The whole data set was split into 10 folds. KNN was run for 10 times with each fold selected as the testing data and the others the training data. The resulted EEPDO will be stored in a vector to be used for the next step.

The KNN regression can be conducted using the build-in KNN algorithm provided by R Package class version 7.3-14.

IV. Experiment

The highway segment of US 49 in Harrison County of Mississippi of a total length of 21 miles was selected as the study area. This roadway segment was selected because it has comparatively high crash records of all crash types. The crash data were derived from the Safety Analysis Management System (SAMS) of MDOT. It contained 641,247 crash claim records with 279 attributes dated from Calendar Year 2011 to 2013. The data have multiple scores including Fatality Analysis Reporting System and Mississippi Department of Public Safety. In the data cleaning process, redundant records were eliminated and study variables were selected. The traffic volume was Annual Average Daily Traffic (AADT) by direction retrieved from Traffic Volume Maps of MDOT for the time that can match the crash data. The road geometry was collected from Google Maps through manual labeling.

According to the raw data, data set "Site" was generated storing the basic observation instances of the sites along the studied roadway. Each site was initialed as an n-dimensional vector (a data table with n attributes), containing attributes of a unique ID, coordinates and location description of starting and ending points. The sites are determined such that the road geometries within each site are homogeneous. The GPS coordinates were used to denote the location of the starting and ending points of the sites. The lengths of sites are normally more than 0.5 miles with an average of 0.73 miles. Separated by direction and by year, the derived site dataset contains 174 instances in total.

The crashes occurred on the intersections are eliminated because the focus of this study is on roadway segments. The crash counts by year (2011, 2012, 2013), by direction (North Bond and South Bond), by severity level and by collision types of each instance were mapped to the sites. The Observed Expected Property Damage Only (OEPDO) was calculated according to the crash-counts-by-injury-severity and weight rules specified by the Traffic Division of MDOT. Of the counted collisions, collision type-1 (Rear end slow or stop, 521 counts in 3 years), type-7 (Sideswipe, 374 counts in 3 years) and type-8 (Angle, 280 counts in 3 years) were found the most frequent in the study area and were extracted for future analysis.

In this study, the NB regression was conducted through a foreign package in R. The resulted APM is shown in (11). Using (4) EEPDOs (predicted value of OEPDO) of all sites were calculated.

$$EEPDO_i = \exp(1.7157 + 0.9577 * segment\ length + 2.6391 * AADT_{perLane}) \quad (11)$$

The ultimate K value of KNN was chosen such that it has the smallest MSE. The MSE output in Fig. 1 shows that K = 4 has the smallest MSE. Therefore, 4 was selected to be the ultimate k.

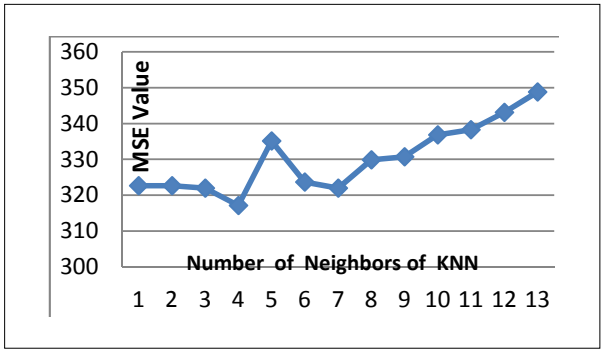

Fig. 1. KNN performance with different numbers of neighbors

In order to evaluate the predication accuracy of the APMs of NB and KNN, the MSE criterion was used to estimate the model performance. Fig. 2 shows the MSE values of two models across the number of neighbors of KNN. Since the NB model was not depend on the number of neighbors of KNN, the graph was a horizontal line with MSE = 866.33. When K = 4, KNN has a smallest MSE of 317.1, which is smaller than NB. The comparison of MSEs indicted that KNN outperform NB in predicting EEPDO.

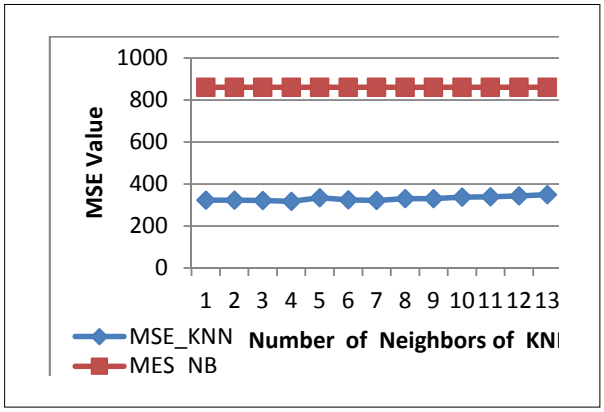

Fig. 2. Prediction performance of KNNN and NB

V. Conclusion

This paper analyzed the mathematical assumptions of NB and KNN as APMs and assessed their predictive accuracies. KNN, a nonparametric predictor, was demonstrated outperform NB in EEPDO perditions. The findings suggest that nonparametric techniques from machine learning present a promising approach for highway HSID. In theory, a nonparametric procedure focuses on the similarities among the attributes of the instances and requires no constraint on the raw data, which makes it an appropriate tool for modeling crash data bearing the intrinsic characteristics of rear and random.

In the future more studies are expected on nonparametric approaches on crash data analysis, especially on testing different road types and how to include priors to crash causation factors such as driver's condition, weather and population characteristics etc that are difficult for statistical APM models to simulate.

Acknowledgment

The project received research funding support from the Institute for Multimodal Transportation (IMTrans) at Jackson State University. Traffic engineers Christopher Kimbrell, Jim Willis, Jessica Dilley, Sammy Holcomb, James Sullivan, and Wes Dean at the Mississippi DOT are thanked for providing data supports to the study

References

[1] AASHTO. "Highway Safety Manual". American Association of State Highway and Transportation Officials, ISBN: 978-1-56051-477-0, Washington DC, 2010.

[2] Khisty, C. J. and B. K. Lall. "Transportation Engineering, an Introduction". Prentice Hall PTR, ISBN 0-13-033560-6, 1980.

[3] Hauer, E. and B. Persaud "Problem of Identifying Hazardous Locations Using Accident Data". Transportation Research Record No. 975, Journal of TRB, National Research Council, Washington, DC, 1984, pp 36-43.

[4] Persaud, B., C. Lyon, and T. Nguyen. "Empirical Bayes Procedure for Ranking sites for Safety Investigation by Potential for Safety Improvement". Transportation Research Record No. 1665, Journal of TRB, National Research Council, Washington, DC, 1999, pp 7-12.

[5] Elvik, R. The predictive validity of empirical Bayes estimates of road safety. Accident Analysis & Prevention, Vol. 40, Issue 6. ISSN: 0001-4575 2008, pp 1964-1969.

[6] Miaou, S-P and J. J. Song. "Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion, and spatial dependence". Accident Analysis and Prevention, Vol. 37 Issue 4, 2005, pp 699–720.

[7] Nelder, J. and R. Wedderburn. "Generalized Linear Models". Journal of the Royal Statistical Society. Series A (General) (Blackwell Publishing), Vol. 135 (3): pp 370–384.

[8] Lord, D. and F. Mannering. "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives". Transportation Research Part A, Vol. 44, Issue 5, 2010, pp 291-305.

[9] Malyshkina, N.V., F. L. Mannering, and A. P. Tarko. "Markov switching negative binomial models: an application to vehicle accident frequencies", Accident

Analysis & Prevention, Volume 41, Issue 2, March 2009, pp 217–226.
[10] Wang, C., M. A. Quddus, and S. G. Ison. "Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model". Accident Analysis and Prevention, Vol. 43, Issue 6, 2011, pp 1979-1990.
[11] Mitchell, T. M. "Machine Learning" McGraw-Hill Science/Engineering/Math, ISBN: 0070428077, 1997.
[12] Cover, M., P.T. Hart. "Nearest Neighbor Pattern Classification". IEEE Transactions of Information Theory, Vol. IT-13, No.1, January 1967, pp 21-27.
[13] Bi, C., X. Ma, Y. Zhang, and Y. Wang "Developing A Cluster-based Algorithm for Collision Hotspot Identification". Proceeding of CICTP 2014: Safe, Smart, and Sustainable Multimodal Transportation, July 4-7, 2014, Changsha, China, pp 2381-2395.
[14] Bandyopadhyaya, R. and S. Mitra. "Fuzzy Cluster–Based Method of Hotspot Detection with Limited Information". Journal of Transportation Safety & Security. Volume 7, Issue 4, 2015, pp 307-323.
[15] Lord, D. and B-J Park. "Negative Binomial Regression Models and Estimation Methods", https://www.icpsr.umich.edu/CrimeStat/files/CrimeStatAppendix.D.pdf, accessed in 2015.
[16] Lord, D. Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accident Analysis & Prevention, Vol. 38, No. 4, 2006, pp. 751-766.
[17] Sawalha, Z. and T. Sayed, "Statistical Issues in Traffic Accident Modeling". Annual Meeting CD-ROM. 82nd Annual Meeting of the Transportation Research Board, Washington DC, January, 2003.
[18] Hauer, E. "Statistical test of a difference between expected accident frequencies". Transportation Research Record No. 1542. The Journal of TRB, National Research Council, Washington, DC, 1996, pp. 24–29.
[19] Hassanat, A.B., M. A. Abbadi, and G. A. Altarawneh. "Solving the Problem of the K parameter in the KNN Classifier Using an Ensemble Learning Approach", International Journal of Computer Science and Information Security, Vol. 12, No. 8, August 2014, pp 33-39.
[20] Butler, Cary. Presentations and Notes of "Machine Learning", unpublished course materials, Jackson State University, Spring 2016.
[21] Hechenbichler, K. and K. Schliep. "Weighted k-Nearest-Neighbor Techniques and Ordinal Classification". Discussion Paper 399, SFB 386, Ludwig-Maximilians University Munich, 2004.