IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 3, March 2020

Intelligent Transport Management System using Data Mining

Dr Rajiv Suresh Kumar[1], Abhijith P T[2], Nissy T Shery[3], Vismaya Pradeep[4]

Department of Computer Science and Engineering,

JCT College of Engineering and Technology, Pichanur (P.O), Coimbatore[1,2,3,4]

Abstract: Traffic congestion is the one the major issue in the present world. This system aims in developing a real time traffic prediction and detection through social conversations by using the technology of data mining. The aim of the system is to predict appropriate class output to each social conversation, whether it is traffic or non-traffic related content. This system employs m-KNN (Modified k Nearest Neighbour algorithm) as a classification model and PCA (Principle Component Analysis) is used for Feature Extraction. This system gives information about the current road status and helps the user to take a better route in their journey.

Keywords: PCA, m-KNN

I. INTRODUCTION

Nowadays, by the increase of vehicles and population there occurs a big challenge in managing the traffic. This system mainly concentrates on the concept of data mining to classify the informative data and to filter out the raw data. In the present world people mostly communicate and convey messages through tweets, so as this scenario exists we go for traffic management through social tweets. Through this system we convey only the traffic related data to the road users based on the current traffic status, so that, the user can select a better route for their smooth journey. This approach helps to overcome the direct and indirect traffic issues. The system uses m-KNN algorithm to classify whether the data is traffic related or not. It also employees PCA for feature extraction. Traffic search is another aspect of this system which helps the user to enquire about the traffic in a particular region. Reducing traffic congestions helps to decrease the number of accidents and also to reduce the waiting time. This paper also describes about the weakness in the proposed system and future directions.

II. EXISTING SYSTEM

Lately, social networks and media platforms extensively using information for detecting each events i.e., traffic congestion, natural disasters, explosion etc. By using special trigger keywords and implementing binary classifier as SVM we can able to detect earthquake through twitter stream and able to identify positive as well as negative events. Fire and explosion can be also be detected by twitter stream analysis by NLP technique and Naive Bayes (NB) classifier.

Disadvantages: The existing system consists of many challenging issues like event detection from social media from traditional media blogs, emails etc where text are well formatted.The other main matter were dealing with problems of text mining which caused lack of precision in thought or communication of natural language due to high dimensionality of feature space will difficult the featured selection in text mining. SUMs are informal and abbreviated words that contains huge misspelling, grammatical and meaningless information that cannot be handled easily.

III. PROPOSED SYSTEM

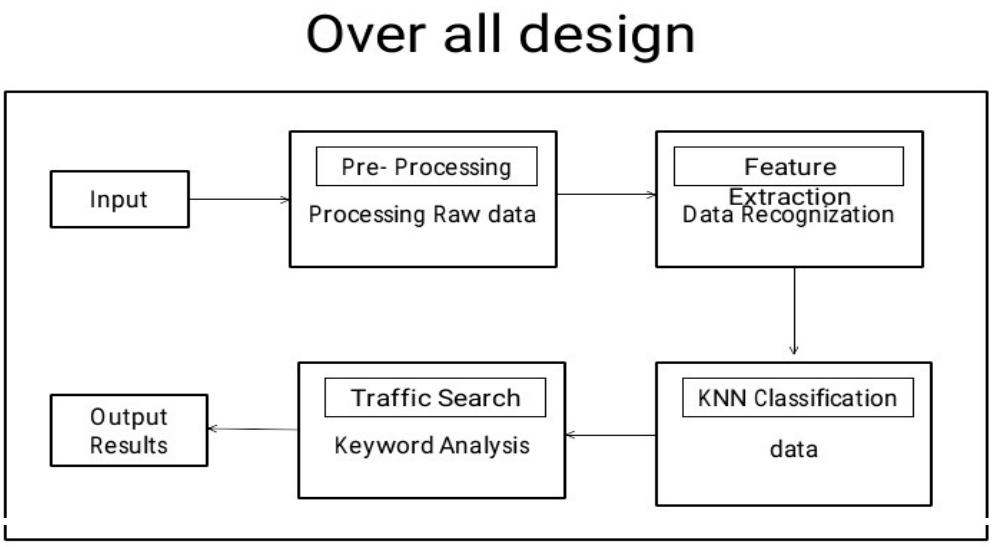
We propose an intelligent system, based on text mining and Naive bayes algorithms, for real-time detection of traffic events from Twitter stream analysis. The system exploits available technologies based on state-of-the-art techniques for text analysis and pattern classification. These technologies and techniques have been analyzed, tuned, adapted, and integrated in order to build the intelligent system. Determining the most effective among different state-of-the-art approaches for text classification. The chosen approach was integrated into the final system and used for the on-the-field real-time detection of traffic events.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

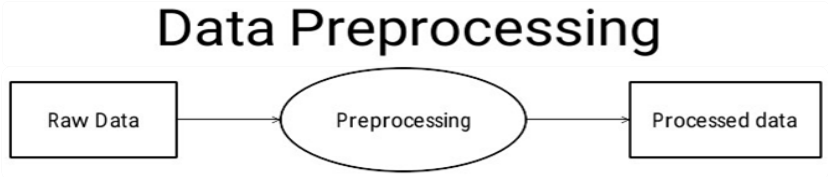
Vol. 9, Issue 3, March 2020



Advantages: Proposed system may approach both binary and multi-class classification problems. As regards binary classification, we consider traffic-related tweets, and tweets not related with traffic. System could work together with other traffic sensors. Intelligent Transportation System monitoring systems for the detection of traffic difficulties, providing a low-cost wide coverage of the road. It performs a multi-class classification, which recognizes non-traffic, traffic due to congestion or crash, and traffic due to external events.

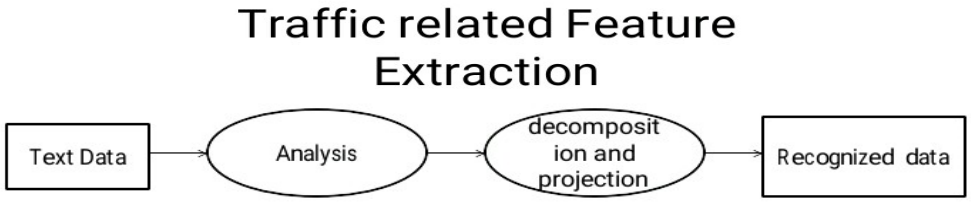
3.1 Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. The process of converting data to something a computer can understand is referred to as pre-processing. One of the major forms of pre-processing is to filter out useless data. In natural language processing, useless words are referred to as stop words such as “the”, “a”, “an”, “in" that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.



3.2 Traffic related Feature Extraction

Feature extraction is the second class of methods for dimension reduction. It creates new attributes (features) using linear combinations of the (original existing) attributes. This function is useful for reducing the dimensionality of high-dimensional data. (i.e. you get less columns).



3.2.1 Text data

Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text. Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interest. Typical text mining tasks include text categorization, text clustering, concept / entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 3, March 2020

3.2.2 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a technique in natural language processing, in particular distributional semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will occur in similar pieces of text (the distributional hypothesis). A matrix containing word counts per paragraph is constructed from a large piece of text and a mathematical technique called Singular Value Decomposition (SVD) is used to reduce the number of rows while preserving the similarity structure among columns. Paragraphs are then compared by taking the cosine of the angle. Values close to 1 represent very similar paragraphs while values close to 0 represent very dissimilar paragraphs.

3.2.3 Data compression

Compression can be either lossy or lossless. Lossless compression reduces bits by identifying and eliminating statistical redundancy. No information is lost in lossless compression. Lossy compression reduces bits by removing unnecessary or less important information. The process of reducing the size of a data file is often referred to as data compression. Compression is useful because it reduces resources required to store and transmit data. Data compression is subject to a space–time complexity trade-off.

3.2.4 Data decomposition and projection

Decomposition is a forecasting technique that separates or decomposes historical data into different components and uses them to create a forecast that is more accurate than a simple trend line. By forecasting each component separately before combining them, you can assess the importance of each and emphasize or discount them according to changing market or economic conditions.

3.2.5 Pattern recognition

Pattern recognition is the automated recognition of patterns and regularities in data. Pattern recognition is closely related to artificial intelligence and machine learning, together with applications such as data mining and knowledge discovery in databases (KDD), and is often used interchangeably with these terms. Machine learning and artificial intelligence is one approach to pattern recognition, By far, the most famous dimension reduction approach is principal component regression. Principal Component Analysis (PCA) is a feature extraction method that uses orthogonal linear projections to capture the underlying variance of the data. PCA can be viewed as a special scoring method under the SVD algorithm. It produces projections that are scaled with the data variance. Projections of this type are sometimes preferable in feature extraction to the standard non-scaled SVD projections.

3.3 Traffic Data Classification

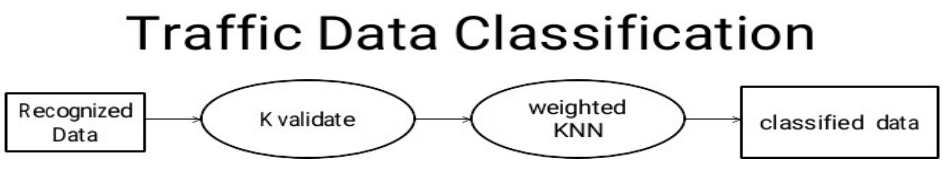
Traffic data is classified using the classification algorithm named Modified K Nearest Neighbor Algorithm (m-KNN). The main idea of the presented method is assigning the class label of the queried instance into K validated data training points. In other hand, first, the validity of all data samples in the train set is computed. Then, a weighted KNN is performed on any test samples. The following shows the pseudo code of the MKNN algorithm.

Pseudo-code of the MKNN Algorithm:

Output label := MKNN (train set , test sample)
Begin
For i: = 1 to train_size
Validity(i) := Compute Validity of i-th sample;
End for;
Output_label:=Weighted_KNN(Validity,test_sample);
Return Output_label ;
End.

In the MKNN algorithm, every training sample must be validated at the first step. The validity of each point is computed according to its neighbors. To validate a sample point in the train set, the H nearest neighbors of the point is considered. Among the H nearest neighbors of a train sample x, validity(x) counts the number of points with the same label to the label of x. Eq. 1 is the formula which is proposed to compute the validity of every points in train set.

$$Validity(x) = \frac{1}{H}\sum_{i=1}^{H} S(lbl(x), lbl(N_i(x))) \quad (1)$$



DOI 10.17148/IJARCCE.2020.9304

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

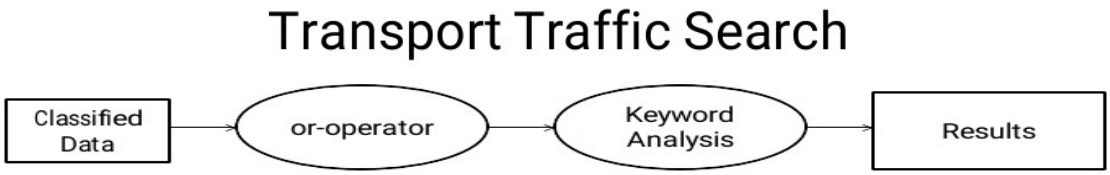
Vol. 9, Issue 3, March 2020

3.4 Transport Traffic Search

The first dataset consists of tweets belonging to two possible classes, namely, i) road traffic-related tweets (class : traffic), and ii) tweets not related with road traffic (class : non-traffic). The tweets were fetched in a time span of about four hours from the same geographic area. First, we fetched candidate tweets for traffic class by using the following search criteria:

— time and date of posting: tweets belong to a time span of four evening hours of two weekend days;

— keywords contained in the text of the tweet: we apply the or-operator on the set of keywords S, composed by the three most frequently used traffic-related keywords, S = {“trafic”(traffic), “rush”(queue), “incident”(crash)} , with the aim of selecting tweets containing at least one of the above keywords.



IV. CONCLUSION AND FUTURE ENHANCEMENTS

System employed m-KNN (Modified k Nearest Neighbour algorithm) as a classification paradigm and PCA (Principle Component Analysis) is used for Feature Extraction. The major component in modern smart transportation systems is Road traffic Prognosis. This method which exceeds the performance of m-kNN method employs a kind of preprocessing on train data that includes a new value named Validity to train samples which cause to more information about the situation of training data samples in the feature space. Thus applying the weighted KNN which employs validity yields to more robust classification rather than simple KNN method, efficiently.

Inorder to identifying traffic patterns, generating a real time information system, developing travel speed calculation model, and investigating parking decisions the descriptive mining methods have aided for these. Predictive data mining techniques used in inferring network topology, finding traffic bottlenecks, solving the multi-objective location inventory problem, constructing two data reduction algorithms and predicting short-term traffic flow in heterogeneous conditions. The main challenging is increasing the potential of the evolutionary model .An effective evolutionary approach without drawbacks will certainly help in developing enhanced ITS.

REFERENCES

[1]. Z. Diao et al., ``A hybrid model for short-term trafc volume prediction in massive transportation systems," IEEE Trans. Intell. Transp. Syst., vol. 20, no. 3, pp. 935946, Mar. 2019.
[2]. (Mar. 2015). TomTom. Accessed: Oct. 11, 2018. [Online]. Available: https://corporate.tomtom.com/news-releases.
[3]. S. Anand, P. Padmanabham, A. Govardhan, and R. H. Kulkarni, ``An extensive review on data mining methods and clustering models for intelligent transportation system," J. Intell. Syst., vol. 27, no. 2, pp. 263273, 2018.
[4]. K.Miller, M.Miller, M.Moran, & B.Dai, ``Data management life cycle," Texas A&M Transp. Inst., College Station, TX, USA, Tech. Rep.1, Mar. 2018.
[5]. J. Raj, H. Bahuleyan, and L. D. Vanajakshi, ``Application of data mining techniques for trafc density estimation and prediction," Transp. Res. Procedia, vol. 17, pp. 321330, Dec. 2016.
[6]. S. Sundaram, S. S. Kumar, and M. D. Shree, ``Hierarchical clustering technique for trafc signal decision support," Int. J. Innov. Sci., Eng. Technol., vol. 2, no. 6, pp. 7282, Jun. 2015.
[7]. K. Kumara, M. Paridab, and V. Katiyar, ``Short term traffic ow prediction for a non urban highway using articial neural network," in Proc. 2nd Conf. Transp. Res. Group India, Agra, India, 2013, pp. 755764.
[8]. D. Schrank, B. Eisele, and T. Lomax, ``TTI's 2012 urban mobility report powered by INRIX trafc data," Texas A&M Transp. Inst. and Texas A&M Univ. Syst., Texas, TX, USA, Tech. Rep. 1, 2012.
[9]. J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, ``Datadriven intelligent transportation systems: A survey," IEEE Trans. Intell. Transp. Syst., vol. 12, no. 4, pp. 16241639, Dec. 2011.
[10]. J. Lopes, J. Bento, E. Huang, C. Antoniou, and M. Ben-Akiva, ``Traffic and mobility data collection for real-time applications," in Proc. 13th Int. IEEE Annu. Conf. Intell. Transp. Syst., Madeira, Portugal, Sep. 2010, pp. 216223.