Multi Output Gaussian Processes for Data Imputation using KNA

[1] Dr. M.Manimekalai, [2] A. Kiruthika

[1]Professor, Director and Head, [2]Research Scholar
[1,2]Department of Computer Science
[1,2]Shrimati Indira Gandhi College, Tiruchirappalli, Tamil Nadu, India-620002

Abstract : These days transportation systems rely increasingly on the convenience and accuracy of traffic detector data to monitor traffic operational conditions and assess system performance. Quality traffic data is an essential for traffic related researches and developing transport applications. There are several techniques to detect the missing data ranging from simply deleting the data to using complex algorithms to impute missing data, traditional methods such as deletion and single imputation and advanced methods such as multiple imputation, model-based procedures, and machine learning techniques. . In this paper, a unique method for missing traffic data imputation is proposed using KNA optimized by a combination of KNN algorithm and SVM. In this method, KNA is the basic algorithm and the parameters of KNA are optimized. Firstly, the patterns of missing traffic data are analysed and the demonstration of missing traffic data is given using matrix-based data structure. Next, traffic data from municipal areas are used to assess spatial-temporal association of the traffic data to resolve**.**

***IndexTerms* - KNA, SVM, KNN, Data mining.**

I. INTRODUCTION

Wireless networks have become increasingly popular in the communication industry. These networks provide mobile users with ubiquitous computing capability and information access regardless of the users' location. There are currently two variations of mobile wireless networks: infrastructure and infrastructure less networks. The infrastructure networks have fixed and wired gateways or the fixed Base-Stations which are connected to other Base-Stations through wires. Each node is within the range of a Base-Station. A "Hand-off" occurs as mobile host travels out of range of one Base-Station and into the range of another and thus, mobile host is able to continue communication seamlessly throughout the network. Example applications of this type include wireless local area networks and Mobile Phone. On considering the location service in a mobile ad hoc network, each node needs to maintain its location information by frequently updating its location with its neighbor nodes which is called as neighborhood update and periodically updating its location information in its network which is called location update. The operation costs in location updates and performance losses for the target application because of the location update has not properly done is a major issue. This may because the network can't be utilized in an effective way.

II. LITERATURE SURVEY

García-Laencina et al. [7, 8] analysed and compared various pattern classi_cation techniques to handle missing data. They presented a top-down pattern classi_cation _owchart, which categorised the various missing data approaches into four groups. They emphasized machine-based solutions and highlighted the advantages and disadvantages thereof. Subsequently, Nishanth and Ravi [7,9] proposed a machine learning technique (probabilistic neural network) which produced ef_cient results when compared to mean, K-Nearest Neighbour (K-NN), Hot Deck (HD) and a decision tree technique. Gómez-Carracedo et al. [7,10] studied air quality data and found that multiple imputation produced more variable results when compared to single imputation methods. Galán et al. [7,11] used genetic algorithms to impute missing data in the knowledge and skills domain. Wang and Chaib-draa [7,12] used an online Bayesian framework incorporating Gaussian Process Regression for surface temperature analysis. The authors concluded that their proposed technique outperforms other Gaussian process techniques such as sparse pseudo-input Gaussian process (SPGP) and sparse spectrum Gaussian process (SSGP).

Finch [7,13] compared the performance of three techniques for imputing missing data for surveys and questionnaires. Multiple Imputation for continuous data (MI), multiple imputation for categorical data (MIC) and stochastic regression imputation (SRI) were compared. It was found that MI or SRI produced less bias than MIC and hence was preferred to MIC. Earlier, Blend and Marwala [7,14] compared an auto-associative neural network (AANN), a neuro-fuzzy (NF) system and a hybrid AANN/NF system in their analysis of Human Immunode_ciency Virus (HIV) and Acquired Immunode_- ciency Syndrome (AIDS) data. It was found that the AANN outperformed the NF system by an average of approximately 6%, while the hybrid method was approximately 16% more accurate than the standalone AANN or NF systems. However; the hybrid system was 50% less computationally ef_cient. Dauwels et al. [7,15] presented an innovative tensorbased imputation method based on canonical polyadic (CP) decomposition which they compared to mean imputation, regression imputation and K-NN. Their proposed method was assessed with medical questionnaires and the results showed that the imputation accuracy improved. Tensor based imputation methods are also widely used methods in traf_c information systems and road sciences and is well documented in literature [7],[16]_[18]. Tensor decomposition techniques are also used in psychology, chemometrics, signal processing, bioinformatics, neuroscience, web mining and computer vision [7,19].

III. METHODOLOGY

These days, most of imputation methods are assessed the missing traffic values by using spatial-temporal information as much as possible; it ignores an essential fact that spatial-temporal information of the traffic missing data is often imperfect and unavailable. Furthermore, most of the existing methods are demonstrated by traffic data from freeway, and their applicability to urban road data needs to be further verified. In this paper, a unique method for missing traffic data imputation is proposed using KNA optimized by a combination of KNN algorithm and SVM. In this method, KNA is the basic algorithm and the parameters of KNA are optimized. Firstly, the patterns of missing traffic data are analysed and the demonstration of missing traffic data is given using matrix-based data structure. Next, traffic data from municipal areas are used to assess spatial-temporal association of the traffic data for the resolve of the proposed method.

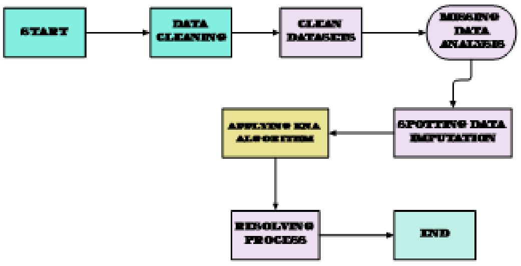


Figure1: system architecture

The proposed system divided into five modules they are home, Dataset analyse Traffic Data, Spotting Missing Data, resolving the missing data.

Imputation is the method of interchanging missing data with substituted values. When substituting for a data point, it is known as "entity imputation"; while exchanging for an element of a data point, it is known as "item imputation". There are three problems that missing data causes: missing data can introduce a considerable amount of preference, make the management and analysis of the data more difficult, and create decreases in efficiency.

A data set is a collection of related, discrete items of related data that may be accessed individually or in combination or managed as a whole entity. A data set is organized into some type of data structure. The database itself can be considered a data set, as can bodies of data within it related to a particular type of information, such as sales data for a particular corporate department.

Traffic information management and analysis systems are currently suffering from poor quality data and missing data. Missing data brings great troubles for further utilization of traffic systems. If any data is missing or poor quality, further process of analysis must be degraded. Imputation methods are used to reduce the impact of incomplete data on utilization.

Missing traffic data can be affected in two categories. First one is loss of data at certain locations and time periods. Complete data is important for analysis in transportation modelling and prediction. For example if speed or volume of traffic data are missing in heavy congested road at peak hours, the vehicle emission level will be under estimated. Second one is statistical information loss i.e primary assumptions of statistical methods used in an imputation process are violated by missing traffic patterns which is giving optimal solutions

The KNA is an algorithm used in data imputation resolving process by simple classification method. The algorithm compared to datasets by feature similarity this means that the new point is assigned a value based on exactly how closely it be similar to the point in the training set. This can be useful in how to handle data and making predictions around the missing data.

Based on prediction the algorithm selects a set of data that describe the decision boundary between classes. KNA is known for excellent classification performance, though it is arguable whether support vectors could be effectively used in communication of medical knowledge to domain experts.

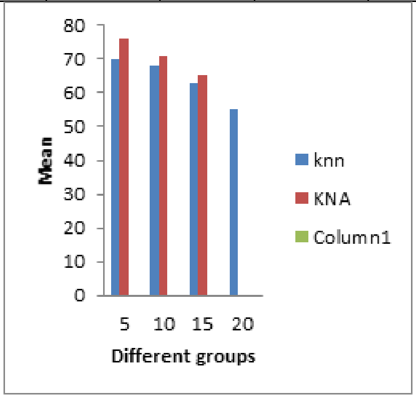
The data sets containing missing values can be processed using an KNA. This is typically accomplished by one of the two means namely ignoring missing data (either by discarding examples with a missing attribute value or discarding an attribute that has missing values), or using a process generally referred to as imputation through, by which a value is generated for the attribute.

IV. RESULT AND DISCUSSION

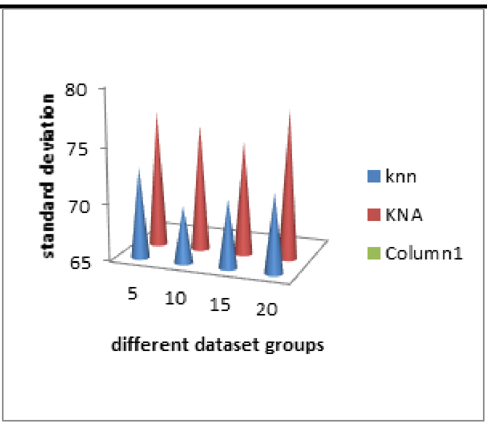
A dataset comprising of 5000 records with 5 factors has been taken for analysis. The test dataset is set up with certain information's missing. The missing rate changes as 2, 5, 10, 15 and 20. KNA calculation is represented with various preparing information and its relating class. The gathering size additionally contrasts as 3, 6 and 9. Each gathering is independently appointed with mean, middle and standard deviation. The outcomes are appeared in the below table.

Table 1: The missing rate changes as 2, 5, 10, 15 and 20

Percentage of imputation	Mean		Median		Standard Deviation	
	Knn	KNA	Knn	KNA	Knn	KNA
5	70	76	73	77	73	77
10	68	71	70	76	70	76
15	63	65	71	75	71	75
20	55	54	72	78	72	78



The table shows the average percentage of the different groups of training dataset. The outcome displays that median and standard deviation has certain progress over mean substitution. There is also a steady improvement in the percentage of accuracy in case of altered sizes of groups.



IV. CONCLUSION

KNA algorithm is a classifier for grouping up of data. This approaches such us Mean/ Median and Standard Deviation is used to increase the performance of accuracy in missing data imputation. This can be further enhanced by comparing with some other machine learning techniques like KNN, SVM. Conclude that the mean error from mean imputation is less effective and the inaccuracy rates are moderately higher than the supervised algorithms. Also, by mean imputation the correlation of attributes in the data set is not taken into consideration by distorting the distribution and underestimating the standard deviation.

REFERENCES

[1] L. Gong and W. Fan, "Applying travel-time reliability measures in identifying and ranking recurrent freeway bottlenecks at the network level," J. Transp. Eng. A, Syst., vol. 143, no. 8, p. 04017042, 2017.
[2] M. Mauch, A. Skabardonis, and L. Davies, "Validating the costeffectiveness model for California's freeway incident management program," Transp. Res. Circular, vol. 1, no. E-C197, pp. 179–191, 2015.
[3] R. J. A. Little and D. B. Rubin, Statistical Analysis with Missing Data. Hoboken, NJ, USA: Wiley, 2014.
[4] L. M. Collins, J. L. Schafer, and C.-M. Kam, "A comparison of inclusive and restrictive strategies in modern missing data procedures," Psychol. Methods, vol. 6, no. 4, pp. 330–351, 2001.
[5] Y. Wang, K. Henrickson, and J. Ash, "Predictive analysis of probe vehicle data completeness," in Proc. TRB Annu. Meeting, 2016, pp. 1–8.
[6] Filipe Rodrigues , Kristian Henrickson, and Francisco C. Pereira, Member, IEEE Multi-Output Gaussian Processes for Crowdsourced Traffic Data Imputation, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS 1524-9050.
[7] MUHAMMAD S. OSMAN, ADNAN M. ABU-MAHFOUZ , (Senior Member, IEEE), AND PHILIP R. PAGE, A Survey on Data Imputation Techniques: Water Distribution System as a Use Case, Received September 27, 2018, accepted October 8, 2018, date of publication October 22, 2018, date of current version November 14, 2018.
[8] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, ``Pattern classi_cation with missing data: A review,'' Neural Comput. Appl., vol. 19, no. 2, pp. 263_282, 2010.
[9] K. J. Nishanth andV. Ravi, ``Probabilistic neural network based categorical data imputation,'' Neurocomputing, vol. 218, no. 12, pp. 17_25, 2016.
[10] M. P. Gómez-Carracedo, J. M. Andrade, P. López-Mahía, S. Muniategui, and D. Prada, ``A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets,'' Chemo- metrics Intell. Lab. Syst., vol. 134, no. 5, pp. 23_33, 2014.
[11] C. O. Galán, F. S. Lasheras, F. J. de Cos Juez, and A. B. Sánchez, ``Missing data imputation of questionnaires by means of genetic algorithms with different _tness functions,'' J. Comput. Appl. Math., vol. 311, no. 2, pp. 704_717, 2017.
[12] Y. Wang and B. Chaib-Draa, ``An online Bayesian _ltering framework for Gaussian process regression: Application to global surface temperature analysis,'' Expert Syst. Appl., vol. 67, no. 1, pp. 285_295, 2017.
[13] W. H. Finch, ``Imputation methods for missing categorical questionnaire data: A comparison of approaches,'' J. Data Sci., vol. 8, no. 3, pp. 361_378, 2010.
[14] D. Blend and T. Marwala. (2008). ``Comparison of data imputation techniques and their impact.'' [Online]. Available: https:// arxiv.org/abs/0812.1539
[15] J. Dauwels, L. Garg, A. Earnest, and L. K. Pang, ``Tensor factorization for missing data imputation in medical questionnaires,'' in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2012, pp. 2109_2112.
[16] H. Tan et al., ``A tensor-based method for missing traf_c data completion,'' Transp. Res. C, Emerg. Technol., vol. 28, pp. 15_27, Mar. 2013.
[17] B. Ran, H. Tan, Y. Wu, and P. J. Jin, ``Tensor based missing traf_c data completion with spatial-temporal correlation,'' Physica A, Stat. Mech.
Appl., vol. 446, pp. 54_63, Mar. 2016.
[18] H. Tan, Z. Yang, G. Feng,W.Wang, and B. Ran, ``Correlation analysis for tensor-based traf_c data imputation method,'' Procedia Soc. Behav. Sci.,
vol. 96, no. 11, pp. 2611_2620, 2013.
[19] M. Mørup, ``Applications of tensor (multiway array) factorizations and decompositions in data mining,'' Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery, vol. 1, no. 1, pp. 24_40, 2011.