Enhanced Preprocessing Algorithm of Information System for Law Enforcement Using Data mining Techniques

A. Malathi, Ph.D
Assistant Professor
Dept of Computer Science
Government Arts College,
Coimbatore

P. Rajarajeswari, Ph.D
Assistant Professor
Department of Mathematics
Chikkanna Govtt Arts College

ABSTRACT
A data preprocessing is a process of cleaning the data, data integration and data transformation. It intends to reduce some noises and inconsistent data. Data preprocessing is the process of keeping the dataset ready for the process. The results of preprocessing step are later used by data mining algorithms. This paper focus on preprocessing the attributes that are related to crime data and that affects the final output of the mining processes.

General Terms
Preprocessing, Your general terms must be any term which can be used for general classification of the submitted material such as Pattern Recognition, Security, Algorithms et. al.

Keywords
Preprocessing, Information system, Law Enforcement, KNN algorithm, EKNN algorithm and EM-

1. INTRODUCTION
Most of the data collection techniques, like survey studies, field experiments, crime findings, etc., produce huge amount of information, where missing values are inevitable. Moreover, data mining techniques, like clustering and classification [8][9], have been designed to analyze and discover knowledge from data that is complete, that is, with datasets that does not contain missing values. The presence of missing values degrades the performance of these data analysis techniques [4] and has to be handled carefully to achieve accurate results. Generally, the analyst has two options to create a dataset containing no missing values.

(i) To delete or ignore these faulty records with missing values

(ii) Fill the missing value with estimated values

Deleting or ignoring rows with missing values have been proved to be inefficient in certain situations and therefore methods that predict the missing values have gained more attention. In the present research work proposes two methods for handling missing values in the crime datasets. As the performance of these proposed methods affect the clustering and classification performance [2], experimental results were conducted with synthetic dataset and their results are used to determine their efficiency on clustering and classification. The winning algorithm will then be used in the proposed crime analysis framework.

The working of the proposed algorithms and the experimental results conducted to analyze their capability in handling missing values are presented in this paper. The traditional algorithms [3][7] which are improved in the present work are described. This paper presents the general methodology used to handle missing values. The working of the two proposed missing values algorithms are presented.

2. PROPOSED ALGORITHM
The general process of the proposed missing value handling algorithms is given here

Step 1 : Partition database into two groups

a) Group 1 : Instances without missing values

b) Group 2 : Instances with missing values

Step 2 : With Group 2, filter all records that have missing values in the selected crime reporting attributes. All these records have no relevancy with the final result and are removed from the original dataset. This reduced Group is termed as Group 2'.

Step 3 : With Group 2', apply the EKNN-LVQ[17] method or Hybrid EM-Naïve Bayes algorithm to handle missing values in individual attributes. Let the resultant group with estimated values be termed as Group 3.

Step 4 : Combine Group 1 and Group 3 to form the complete data without missing values

Fig 1: General Methodology

2.1. The EKNN-LVQ Algorithm
The steps in the EKNN-LQV [11][12] method are given below.

Step 1 : Read the incomplete dataset

Step 2 : Train SRNG with the scaled Euclidian metric (SEM)

Step 3 : Initialize relevance factor and λ_I uniformly and select relevant neighbours.

Step 4 : For each incomplete pattern, from the selected k neighbours, impute new values.

The advantage of this step is that the process of finding mutual information data [5][6] for identifying relevant values is simplified by using LVQ based SRNG[10]. This when combined with enhanced KNN imputation method increases the accuracy. The experimental results are presented.

2.2. Hybrid Em-Naïve Bayes Imputation Method

The proposed algorithm uses a two-step approach, where the first step clusters the selected attribute into groups and the second step classifies records with unknown values into predetermined classes. The justification for using clustering is as follows: Classes from clusters are more likely to represent the actual real word data and needs only a single attribute value of each record. These values were clustered using EM algorithm [1] and initially 10 clusters were chosen and then the classification process is performed using naïve bayes [15] classifier. The naïve bayes classifier is chosen because of simplicity during implementation. The model is referred as EM-Naïve Model in this paper.

The general steps of the proposed algorithm are given below.

Step 1: Let X be the incomplete dataset. Partition X into Y_1 and Y_2 in such a way that Y_1 has only complete data and Y_2 has data with missingness.

Step 2: Use simple EM-algorithm to cluster the attribute of Y_1 dataset with parameters set to 100 iterations, minimal standard deviation 1.0E-6 with seed 100.

Step 3: Use Y_1 dataset to train the classifier. The classes are assigned to closely to estimate value to avoid large gaps.

Step 4: Each class is converted to an estimated value.

Step 5: The final step was to predict the population class unknown attributes using naïve bayes classifier.

3. DATASET

The original data set was collected for 10 years from Commissioner Office, Coimbatore, Tamil Nadu. . The dataset contains the following categories.

- Rape
- Molestation
- Kidnapping And Abduction
- Sexual Harassment

This dataset was used for the implementation, but the best

algorithms are chosen by using the synthesized data set.

4. PERFORMANCE EVALUATION

This section presents the results obtained while using the synthetic dataset for evaluation. The experiments were conducted with the objective of testing the proposed models in their efficiency in handling missing values. The performance evaluation is performed in three stages. The first stage uses the metrics NRMSE and execution time to analyze the performance of the proposed system with varying data size and missing percentage. The second stage uses the metrics classification accuracy to analyze the impact of missing handling prediction on classification. The third stage, through the use of silhouette measure, analyzes the impact of missing value prediction on clustering.

For this purpose, the dataset was created by varying only the data size and missing percentage parameters[16]. The data size was varied with values ranging from 1000 to 5000 in steps of 1000 and the missing percentage was varied with values ranging from 10 to 40% in steps of 10. .

5. EVALUATION OF MISSING VALUE PREDICTION

The results obtained by the two enhanced algorithms, EKNN-LVQ [13][14] and EM-Naïve is compared with their traditional counterparts KNN and EM algorithms, are presented in this section.

5.1 Normalized Root Mean Square Error (NRMSE)

The performance of the two proposed algorithms with respect to Normalized Root Mean Square Error (NRMSE) is presented in this section

From the results obtained it can be seen that both the proposed algorithms, EKNN-LVQ [17] and EM-Naïve approaches to handle missing data is efficient in terms of NRMSE. Both algorithms are efficient when compared with the KNN and EM traditional algorithms. This is evident from the low NRMSE values (near to zero) obtained. The average NRMSE was calculated for each dataset.

The results further indicate that both the algorithms scale well with both small-sized and large-sized datasets. The dataset size and NRMSE values are inversely proportional to each other. That is, as the dataset size increases, the NRMSE value decreases, which indicate the prediction of missing values, are becoming more accurate. A similar trend was observed with missing value percentage also.

Table1. Normalized mean square error

Datasets	% of Missingness	KDD	EKNN-LVQ	EM	EM-NAÏVE
1000	10	0.6691	0.6141	0.6673	0.6140
	20	0.6675	0.6211	0.6657	0.6209
	30	0.6688	0.6306	0.6670	0.6310
	40	0.6683	0.6318	0.6665	0.6320
2000	10	0.6714	0.5832	0.6696	0.5833
	20	0.8095	0.6377	0.8077	0.6378
	30	0.8009	0.6312	0.7991	0.6311
	40	0.6876	0.6541	0.6858	0.6544
3000	10	0.5415	0.5214	0.5397	0.5221
	20	0.5548	0.5334	0.5530	0.5330
	30	0.6600	0.5940	0.6582	0.5936
	40	0.6775	0.5891	0.6757	0.5903
4000	10	0.5251	0.4569	0.5233	0.4582
	20	0.5726	0.5519	0.5708	0.5518
	30	0.5410	0.4217	0.5392	0.4216
	40	0.5573	0.4312	0.5555	0.4314
5000	10	0.5451	0.4786	0.5433	0.4782
	20	0.5049	0.4210	0.5031	0.4280
	30	0.5103	0.4318	0.5085	0.4334
	40	0.5282	0.4533	0.5264	0.4513

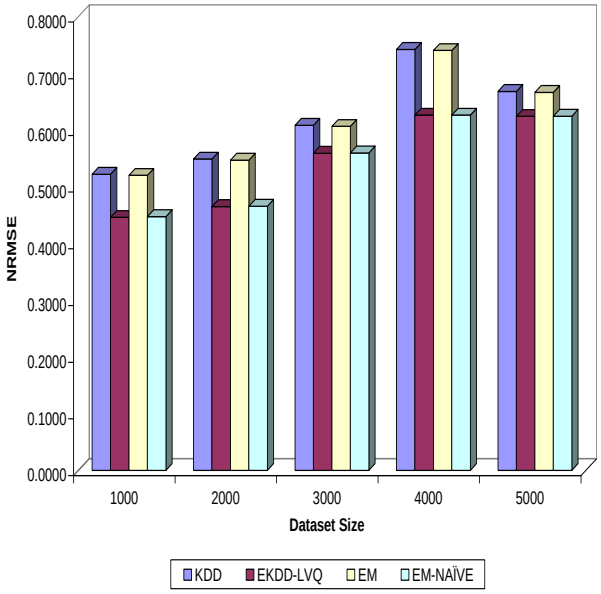


Fig.2 : Average NRMSE Of Traditional And Proposed Algorithms

From Fig 2, it is again clear that the proposed models are improved version of their traditional counterparts. The EKNN-LVQ algorithm showed on average 12% efficiency gain over KNN algorithm, while the EM-Naïve algorithm showed 11.66% efficiency gain over EM algorithm. This shows that the performance of both the algorithms in treating missing values is more or less the same, with only 2.86% accuracy gain showed by EM-Naïve algorithm over EKNN-LVQ algorithm.

The main objective of this experiment is to pick out an algorithm that works predicts missing values in crime data efficiently. From the results, it could be inferred that the performance of the proposed algorithms are more or less similar. In order to meet the objective of this experiment, it was decided to analyze their impact on classification and clustering performance. The next two sections describe these results.

5.2 Impact on Classification Performance

Table 2 shows the classification accuracy of the proposed EKNN-LVQ and EM-Naïve models and compares the results with traditional KDD and EM model. The table projects obtained results of the experiments while varying both the dataset size and missing value percentage. Classification was performed on Crime Type.

The results prove that the algorithms perform better when supplied with more data. That is the accuracy of the classifier increases when the dataset set size increases. This is evident from the increasing tendency of accuracy obtained while varying the data size from 1000 to 5000. The minimum accuracy obtained by the classifier with dataset 1000 is 84.09 and 85.62 while maximum accuracy is 90.64 and 92.33. This shows that the scalability property of both algorithms is well maintained.

To analyze the performance of the algorithm against each other the average value was calculated for each dataset size and the results are projected in fig 3.

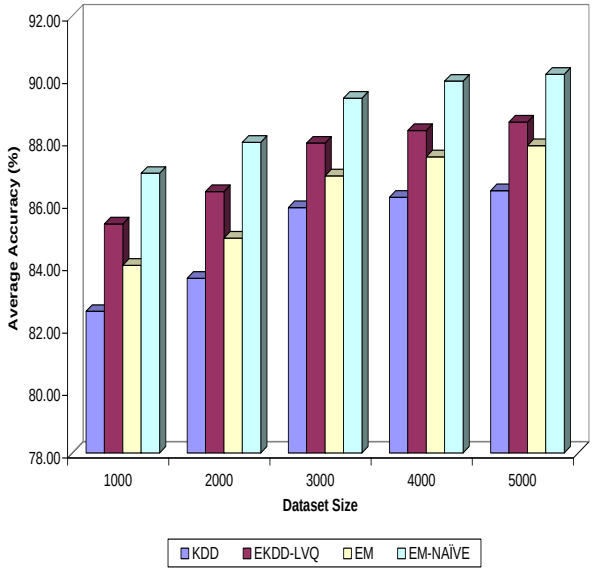


Fig 3: Average Classification Accuracy

From the fig 3, it is evident that both the proposed algorithms have increase in efficiency with respect to classification accuracy when compared with to traditional algorithms. The EKNN-LVQ algorithm showed an average.

Table 2. Classification Accuracy

Dataset Size	% of Missingness	KDD	EKDD-LVQ	EM	EM-NAÏVE
1000	10	81.86	84.09	82.65	85.62
	20	82.06	84.82	83.46	86.14
	30	82.62	85.57	84.68	87.88
	40	83.60	86.87	85.26	88.21
2000	10	82.92	85.42	83.80	86.34
	20	83.14	85.71	84.23	87.31
	30	83.87	86.77	85.23	89.23
	40	84.45	87.60	86.27	88.93
3000	10	84.00	86.71	85.01	87.22
	20	84.95	86.99	85.97	88.93
	30	86.19	88.17	87.61	89.77
	40	88.30	89.84	88.92	91.54
4000	10	84.61	86.74	85.73	87.61
	20	85.11	87.67	86.72	89.46
	30	86.38	88.69	88.22	90.92
	40	88.67	90.23	89.29	91.66
5000	10	84.71	86.89	85.79	88.32
	20	85.55	87.73	87.17	88.95
	30	86.61	89.14	88.42	90.93
	40	88.75	90.64	90.00	92.33

Efficiency gain of 2.75% while the EM-Naïve algorithm showed 2.97% efficiency gain in accuracy when compared with KDD and EM algorithms respectively. While comparing between the two proposed models, it is evident that the EM-Naïve algorithm is efficient than the EKDD-LVQ algorithm. The EM-Naïve algorithm on average shows an average efficiency gain of 7.72%.

From the various results, it can be seen that the EM-Naïve algorithm produces better classification results. This motivated the researcher on the decision to use EKNN-LVQ algorithm for classification purposes during the design of crime analysis framework. As classification plays an important role during prediction of data, it was decided to use this for imputing missing values in the prediction of number of crimes reported.

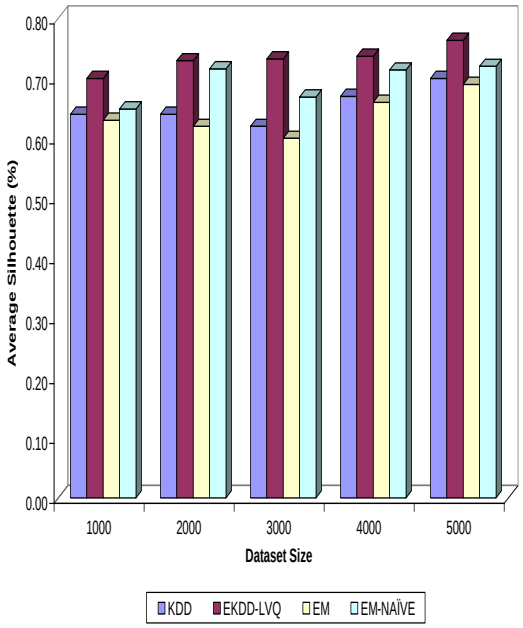
5.3 Impact of clustering Performance

The performance of the two proposed algorithms with respect to clustering performance in terms of Silhouette measure is presented in Table 4.3. Clustering was based on the crime type attribute.

Table 3. Silhouette Measure

Dataset Size	% of Missingness	KDD	EKDD-LVQ	EM	EM-NAÏVE
1000	10	0.64	0.70	0.65	0.66
	20	0.63	0.70	0.64	0.65
	30	0.63	0.70	0.64	0.65
	40	0.62	0.70	0.63	0.64
2000	10	0.63	0.74	0.64	0.72
	20	0.63	0.73	0.64	0.71
	30	0.63	0.73	0.64	0.72
	40	0.62	0.72	0.64	0.71
3000	10	0.61	0.70	0.63	0.66
	20	0.61	0.71	0.62	0.67
	30	0.61	0.75	0.63	0.66
	40	0.60	0.74	0.62	0.67
4000	10	0.67	0.72	0.68	0.70
	20	0.66	0.73	0.67	0.71
	30	0.66	0.74	0.67	0.71
	40	0.66	0.74	0.66	0.72
5000	10	0.70	0.75	0.71	0.71
	20	0.70	0.76	0.71	0.72
	30	0.68	0.77	0.70	0.72
	40	0.67	0.76	0.69	0.72

As with classification results, the clustering results also show that the performance of the algorithms increases with dataset size and missing value percentage. The clustering efficiency of the EKNN-LVQ algorithm ranged between 0.70 and 0.78 while it was between 0.60 and 0.71 for KNN algorithm. Similarly, the EM-Naïve algorithm produced



classification accuracy in the range 0.64 and 0.73, while EM algorithm produced values in the range 0.59 and 0.70. This proves that the proposed algorithms are enhanced versions of their base algorithms.

To compare the efficiency of the two proposed algorithms, the average Silhouette measure for each dataset size was calculated and was obtained from the above data. The results obtained are projected to analyze the overall performance of the proposed algorithms.

From the above Figure, it is clear that the EKNN-LVQ algorithm performs better than EM-Naives algorithm in terms of clustering efficiency. When compared with traditional KNN and EM algorithm, the EKNN-LVQ showed an average silhouette gain of 10.71% and 7.72% respectively. While comparing the two proposed algorithms, the EKNN-algorithm showed a higher efficiency gain of 27.91%.

This motivated the researcher to use EKNN-LVQ algorithm for clustering crime data during preprocessing. As clustering has more on grouping similar results together, it was decided to use this technique for predicting missing values in the population size attribute of the crime dataset.

Thus, from the various results, it could be concluded that the performance of proposed missing handling procedures produced similar results with respect to NRMSE and speed. However, the effect of predicting missing values using EKNN-LVQ algorithm when combined with clustering produced higher accuracy than EM-Naïve algorithm. When analyzed with classification accuracy, the EM-Naïve algorithm performed better than EKNN-LVQ algorithm. So, as mentioned earlier, it was decided to use EKNN-LVQ algorithms to predict the population size attribute in crime dataset, as it requires clustering of data into states and area before the missing value can be determined. Similarly, it is further decided to use the EM-Naïve algorithm for predicting missing values in the number of crimes reported, as it requires only classifying the missing values.

5. CONCLUSION

This paper focused on improving the missing handling procedures for efficient clustering and classification processes. Two methods, one that enhances the traditional KNN algorithm and another that improves the traditional EM algorithm were used for predicting missing values. A crime synthetic dataset was used to analyze the performance of the proposed algorithms.

Thus, it was observed that both the proposed algorithm showed significant improvement to their traditional algorithms and the results of the two algorithms were consistent and close to each other. In order to determine the best among the two algorithms, the experiment was extended to analyze their efficiency in terms of classification and clustering. The results portrayed the fact that while taking classification into consideration, the hybrid model that improved EM algorithm by combining it with naïve bayes classification performed better than EKNN-LVQ algorithm. Therefore, based on these results, it was decided to use the EM-Naïve algorithm to predict the population size of a city. While taking clustering into consideration, the enhanced KNN algorithm, EKNN-LVQ, showed significant improvement over EM-Naïve and hence will be used to predict values for all attributes that report the number of crimes.

The main objective of this paper is to convert an incomplete dataset with missing values to a complete dataset that can be handled efficiently by the proposed crime data analysis framework. Two important tasks in the proposed framework are classification and clustering operations.

6. ACKNOWLEDGMENTS

This paper is an outcome of a project funded by UGC. We are very thankful to the University Grants Commission, South Eastern Regional Office, Hyderabad.

We are also thankful to The commissioner and his team members for sharing their valuable knowledge in this field.

7. REFERENCES

[1] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood for incomplete data via the EM algorithm. J. R. Stat. Soc., Ser. B, Vol. 39, Pp.1–38.

[2] Duda, R.O., Hart, P.E. and Stork, D.G. (2000) Pattern Classification, Wiley-Interscience, New York.

[3] Hammer, B. and Villmann, T. (2002) Generalized relevance learning vector quantization, Neural Networks, Vol. 15, Issues 8–9, Pp. 1059–1068.

[4] Kapur, J.N. (1994) Measures of Information and their Application, Wiley, New Delhi.

[5] Kohonen, T. (1997) Self-Organizing Maps, vol. 30 of Springer Series in Information Sciences, Springer, Berlin, Heidelberg, 1995, (Second Extended Edition 1997).

[6] Kwak, N. and Choi, C.H. (2002) Input feature selection by mutual information based on Parzen window, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No.12, Pp. 1667–1671.

[7] Meyering, A. and Ritter, H. (1992) Learning 3D-shape-perception with local linear maps, Proceedings of IJCNN'92, Pp. 432.436.

[8] Martínez-Muñoz, G. and Suárez, A. (2004), Using all Data to Generate Decision Tree Ensemble, IEEE Tran. On Systems, Man and Cybernetics—part C: Applications and Review, Vol. 34, No. 4, Pp. 393-397.

[9] Peres, R.T. and Pedreira, C.E. (2009) Generalized Risk Zone: Selecting Observations for Classification, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, No. 7, Pp. 1331-1337.

[10] Pregenzer, M., Pfurtscheller, G. and Flotzinger, D. (1996) Automated feature selection with distinction sensitive learning vector quantization. Neurocomputing, Vol. 11, Pp.19-29.

[11] Sato, A. and Yamada, K. (1998) A formulation of learning vector quantization using a new misclassification measure, Proceedings of Fourteenth International Conference on Pattern Recognition, A. K. Jain, S. Venkatesh, and B. C. Lovell, Eds., IEEE Computer Society, Los Alamitos, CA, USA, Vol. 1, Pp. 322-325.

[12] Sebastiani, F. (2002) Machine learning in automated text categorization, ACM Computing Surveys, Vol. 34, No. 1, Pp. 1- 47.

[13] Villmann, Th. and Hammer, B. (2002) Supervised neural gas for learning vector quantization, Proc. of the 5th German Workshop on Artificial Life (GWAL-5), D. Polani, J. Kim, and T. Martinetz, Eds.,Akademische Verlagsgesellschaft - infix - IOS Press, Berlin, Pp. 9-16.

[14] Villmann, Th., Schleif, F. and Hammer, B. (2006) Comparison of Relevance Learning Vector Quantization with other Metric Adaptive Classification Methods, Neural Networks, Vol. 19, Issue 5, Pp. 610-622

[15] Zhang, H. (2004) The optimality of Naive Bayes, American Association for Artificial Intelligence, Flairs Conference, Pp. 257-319

[16] Adèr, H.J. and Mellenbergh, G.J. (Eds.) (2008) Chapter 13: Missing data, Advising on Research Methods: A consultant's companion, Huizen, The Netherlands: Johannes van Kessel Publishing, Pp. 305-332.

[17] Chen, N., Vieira, A., Ribeiro, B., Duarte, J. and Neves, J (2011) A stable credit rating model based on learning vector quantization, Vol. 15, No. 2/2011, Intelligent Data Analysis, Pp. 237-250