IJCSC International Journal of Computer Science and Communication Vol. 2, No. 2, July-December 2011, pp. 337-343

EFFICIENT KNOWLEDGE DISCOVERY USING DATA CLUSTERING FOR TRANSPORTATION PLANNING

Sai Hanuman A.[1], Anand Sesham[2], Vinaya Babu[3], Govardhan[4] and Padmanabham[5]

[1]Associate Professor of CSE, HOD of MCA, GRIET, Hyderabad, India. *E-mail: a_saihanuman@hotmail.com*
[2]Associate Professor of CSE, MVSREC, Hyderabad, India. *E-mail: sesham_anand@hotmail.com*
[3]Professor of CSE, Director of Admissions, JNTU Hyd, India. *E-mail: dravinayababu@yahoo.com*
[4]Professor of CSE, Principal, JNTUH, Jagitial, A.P., India. *E-mail: govardhan_cse@yahoo.co.in*
[5]Principal, Bharath Engg. College, Hyderabad, India.

ABSTRACT

Data mining is the process of extracting hidden patterns from the given data. With the explosive increase of data every year, data mining is becoming an increasingly important tool to transform this data in to information. In this exploration we attempt to apply various Data clustering techniques to a Home Interview Survey Data related to Transportation planning. Real world databases contain a lot of noisy, missing and inconsistent data because of huge size and often errors occurred during data collection. In order to improve the quality of data mining results, we need to preprocess the data by applying various techniques. In our experimentation we applied preprocessing and clustering techniques on the realistic Transport dataset and could able to draw useful inferences. The suitability of various conventional and advanced data mining techniques on this kind of a dataset is explored and a conclusion is drawn. Also we demonstrated the need of Evolutionary computational techniques for solving these kind of problems.

Keywords: Knowledge Discovery, Prediction, Partitional Clustering, HIS, ECTs, Factor Analysis, PCA.

1. INTRODUCTION

Data mining is the process of extracting hidden patterns from the given data. With the explosive increase of data every year, data mining is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of profiling practices such as marketing, fraud detection and scientific discovery. Data mining can be applied to data sets of any size. The term data mining is often used to apply to the two separate processes of Knowledge Discovery and Prediction. The term Knowledge Discovery from Databases(KDD) [1-2]refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It involves the evaluation and possibly, interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step. KDD [3] refers to the overall process of discovering useful knowledge from data.

The overall process of finding and interpreting patterns from data involves the repeated application of the following steps:

- Learning the application domain
- Choosing the data mining algorithm(s)
- Creating a target dataset
- Data mining
- Data cleaning and preprocessing
- Interpretation
- Data reduction and projection
- Using discovered knowledge
- Choosing the function of data mining

KDD Techniques

There are many different approaches that are classified as KDD techniques. Some of them discussed in [3] are:

- Probabilistic Approach
- The Decision Tree Approach
- Statistical Approach
- Deviation and Trend Analysis
- Classification Approach
- Neural networks
- The Bayesian Approach to KDD
- Hybrid Approach
- Pattern Discovery and Data Cleaning

Before applying any of these techniques the data is to be properly organized and cleaned. Real world databases and data available in files contain a lot of noisy, missing and inconsistent data because of huge size and often errors occurred during data entry. In order to improve the quality of data and also the quality of data mining results, we need to preprocess the data by applying various techniques. There are a number of data pre processing tasks available [4]. They are:

1. Data cleaning
2. Data integration and transformation
3. Data reduction
4. Discretization and concept hierarchy generation

Some of the above approaches have been applied on the target dataset. Details of the entire process of data cleaning for the given data are illustrated in [4].

The main tasks of Data Mining can be classified as Classification, Clustering, Regression and Association rule mining. In this paper we mainly focused on the Data clustering for Knowledge extraction.

2. DATA CLUSTERING

This is like classification but the groups are not predefined, so the algorithm will try to group similar items together. Data clustering is well known as unsupervised classification.

Clustering techniques can be broadly classified as follows [5]

- Hierarchical clustering
- Combinatorial search techniques-based clustering
- Partitional clustering
- Fuzzy clustering
- Density based clustering
- Neural network based clustering and
- Graph theory-based clustering
- Kernel based clustering

In this paper we explored mainly on partitional based clustering algorithms.

3. DATASET DESCRIPTION

In our exploration we applied various partitional based clustering approaches to transportation domain. In this we used a dataset related to a transportation project, named as Home interview Survey(HIS) data. The main purpose of organizing such a large scale Home Interview Surveys was to understand the present day travel patterns and relate these travel patterns to the Socio-Economic characteristics of Trip makers, the type of activity after reaching the end of trip, and to the transportation supply provided in the form of road network, public transportation facilities etc. While so doing, several socio economic and vehicle ownership properties are captured through a set of questions administered on them. It is now proposed to use this extensive data to derive some additional and useful information that may further help in understanding the socio economic characteristics for transportation planning.

3.1 Background of Home Interview Surveys in Travel Demand Modeling

The travel patterns in the form of "Number of Trips" performed by each member in a city, from a identifiable location in the city called "Origin" to another identifiable location called "Destination", together with the trip makers "Socio Economic characteristics", is the primary bed block based on which future predictions of travel are made. This information is used in developing Travel Demand Models that will help in predicting future travel patterns for the horizon year. These predicted travel patterns are the main source of information for identifying, planning, locating, designing, justifying various transportation projects [8]. For calibrating the Demand model, base year travel patterns along with their Attributes are necessary. For this purpose elaborate travel surveys are organized.

The main method of obtaining all these travel attributes from road users is to elicit from them either directly by interviews or by obtaining indirectly by phone call or through written reply by mail or e-mail. The principal methods of intercepting the transport users are either at the "Beginning of trip called Origin end", or at "End of called Destination end", or "En route called Road side survey" of the travel.

If the surveys are organized at Origin end, it is usually the Home; the information collected is termed as "Home Based Survey". If the surveys are organized during the course of travel, they are usually termed as "Road Side", or "Cordon interviews" depending upon weather the location is within study area or at the cordon location. If on other hand the road user is interviewed after reaching the destination, the survey is termed as Destination Based survey. In fact all the techniques are to be employed for capturing complete information on urban travel.

In order to present the findings of these surveys it is necessary to translate the information into Origin Destination matrices stratified by Purpose, Mode, Time of Travel etc, for further applications.

Since there are innumerable number of Origins and Destinations in the study region, it is not possible to describe each trip from their exact place of origin to exact place of destination. Rather, the study area is divided into small "Traffic Analysis Zones (TAZ)" or localities to represent a group of houses or group of activities. All those trips that start anywhere in the Traffic Analysis Zone (TAZ) area, or end anywhere in the area, they are assumed to be originating at or destined to the centroid of the TAZ. Thus all individual homes and activities are aggregated to reasonable number of representative traffic analysis zones. This finite number of representative origins and destinations enables to generate O-D matrices that are reasonably workable, for computations and model development. However care is to be exercised to ensure that zone sizes are not too big as to distort the travel patterns, or nor too small that the secondary data becomes difficult to obtain and predict for future, or becomes not compatible with transport network. In the present study the data was obtained from Hyderabad city Home Interview Surveys, divided into 147 Traffic Analysis Zones or localities.

The main issue is to capture attributes that are relevant, causative and should not contain noise or redundancy. Data mining techniques can help in identifying variables that contain relevant information for use in the models. We now present the data that is normally collected through Home Interview Surveys.

3.2 Home Interview Survey Format

The questionnaire was so designed that all the planning parameters required for transportation planning are captured in a specially designed format by the study group. Details of the format and how the samples were created can be found in [7]. As will be seen, several questions are framed to elicit extensive data that may be useful. In the process there are several questions that may contain redundant or noise information.

4. ALGORITHMS USED

As part of a study being done, we have applied different data mining techniques namely, Factor Analysis, for identifying basic traits in the socio economic data, and Clustering for grouping the urban commuters into homogenous groups for stratifying the urban commuters for modeling travel behavior. For our experimentation we used SPSS 11.0 , Java and WEKA[9].

4.1 Principal Components

The data collected from Home Interview Surveys contains inter correlations among the elements of the socio economic vector variable which makes it difficult to construct and interpret any model from them. To detect these relationships one can use projections along different directions defined by a weighted linear combination of variables, or components that maximize the variance subject to being uncorrelated. Principal component analysis is a useful procedure to determine the minimum number of independent dimensions needed to account for most of the variance in the original test data. They do reveal fewer independent dimensions that are required to define the test domain.

4.2 Factor Analysis

Factor analysis further improves the solutions offered by principal components by rotating components to positions that are most interpretable. With only few variables, it is easier to search interesting spaces manually by rotating the distribution data. In this paper we have employed varimax solution for rotation of axis. For each factor, varimax rotation also yields high loadings for a few variables that will help in understanding basic trait associated with the factor.

4.3 Clustering using K-means Algorithm

Since the present data has mixed variable scales, Simple *k*-Means [6] clustering algorithm, was found appropriate for further analysis. After the initial preprocessing of data, the cleaned and mining ready data file was then used to be applied using these algorithms.

5. EXPERIMENTATION

Table I presents the list of 26 socio economic variables collected from home interview survey that have been standardized to have zero means and unit variances and subsequently converted into a correlation matrix **R**.

Table 1
List of Variables

Variable No.	*Attribute*	*Full Name*
House hold details		
1	MALES	Total number of males
2	FEMALES	Total number of females
3	TOTAL_NO	Total members in household
4	NO_MEM_STU	No. of members in the family studying
5	VO_CY	Total number of cycles in the house
6	TWO_WHLR	Total number of motorcycles in the house
7	VO_CAR	Total number of cars in the house
8	TOTAL_VEHI	Total No. of vehicles in the house
9	MONTH_INCO	Total Monthly income in Rs of house hold
10	MONTH_TRAN	Monthly expenditure Rs on transport in house hold
11	RESIDENCE	Residence Type; Owned or Rented
Personal Details		
12	AGE	Age in years
13	SEX	Sex Male or Female
14	EDUCATION	Education level code 1 to 5
15	INCOME	Income per month Rs
16	OCCU_CODE	Occupation code 1 to 11
17	VO_CODE	Vehicle ownership (person)
18	RAIL_BUS_P	Rail/Bus pass holder

Table Contd…

Table 1 Contd…

Travel details		
19	DISTANCE	Total distance traveled km
20	PURPOSE	Purpose of travel 1 to 6, 11 to 16
21	S1_DISTANCE	Stage distance km
22	S1_MODE	Mode of travel code 1 to 12
23	S1_TRA_T1	Travel time minutes
24	S1_TYPE_PA	Type of parking 1 to 4
25	S1_COST_PA	Cost of parking Rs per trip
26	S1_TRAV_CO	Travel cost Rs

The Eigen values have been obtained as the roots of the characteristic equation

$$|\mathbf{R} - \lambda_i \mathbf{I}| = 0$$

Only those Eigen values which are greater than 1.0 were retained. A scree plot was used to decide as to how many Eigen values to be retained. It is noted that only 6 principal components could explain about 65% of the total variance of the original data.

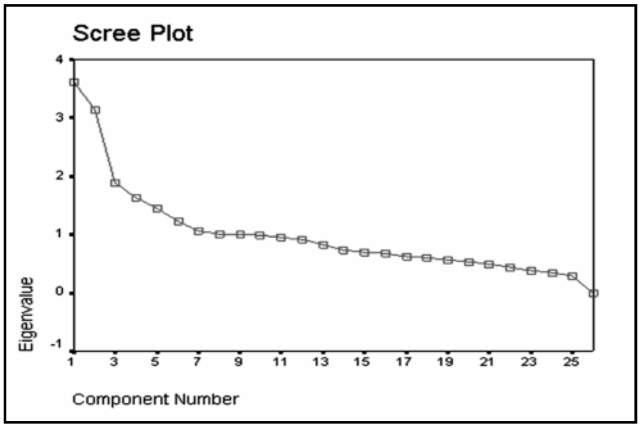


Figure 1: Scree Plot Obtained

Table 2
Eigen Values

Number	*Eigen Value*	*Proportion*	*Cumulative*
1	3.34963	0.13399	0.13399
2	2.88115	0.11525	0.24923
3	2.88115	0.08881	0.33804
4	1.87342	0.07494	0.41298
5	1.41665	0.05667	0.46964
6	1.22786	0.04911	0.51876
7	1.11165	0.04447	0.56322
8	1.05815	0.04233	0.60555
9	1.0025	0.0401	0.64565
10	0.99124	0.03965	0.6853
11	0.91829	0.03673	0.72203
12	0.79481	0.03179	0.75382
13	0.75648	0.03026	0.78408
14	0.70476	0.02819	0.81227
15	0.67287	0.02691	0.83919
16	0.62831	0.02513	0.86432
17	0.60904	0.02436	0.88868
18	0.59162	0.02366	0.91234
19	0.55513	0.02221	0.93455
20	0.48025	0.01921	0.95376

A factor rotation based on Kaisers' criterion was performed. The loadings are so available that some scores have heavy coefficients while other coefficients are less loaded. These coefficients of the Factor Matrix indicate the correlations of the variables with respective factors and provide valuable basis for defining them. As shown in the following Table.

Table 3
Rotated Eigen Vectors

	3.34963	*2.88115*	*2.88115*	*1.87342*	*1.41665*	*1.22786*
Variable No	*V1*	*V2*	*V3*	*V4*	*V5*	*V6*
1	-0.0439	-0.4162	0.0787	-0.1793	-0.1591	-0.1899
2	-0.0735	-0.4284	0.0731	0.0333	0.0932	0.1191
3	-0.0753	-0.5435	0.0977	-0.0962	-0.045	-0.0488
4	-0.0982	-0.3484	0.0855	-0.1535	0.209	-0.0109
5	-0.0774	-0.1055	0.0153	-0.342	-0.3039	0.1696
6	0.0024	-0.1142	-0.0014	-0.2613	-0.3198	0.2048
7	0.2164	-0.1201	0.0414	0.4151	0.0631	0.1194
8	-0.0255	0.089	0.6383	0.0363	-0.0841	0.0669
9	0.2424	-0.2748	0.0538	0.3412	-0.1072	-0.035

Table Contd…

Table 3 Contd…

10	0.2433	-0.2222	0.0474	0.3856	0.013	0.0136
11	-0.0636	0.1971	0.3216	-0.0465	0.0066	0.1168
12	0.1672	0.0588	-0.0539	0.1056	-0.5122	-0.0759
13	-0.2325	-0.0258	-0.0123	0.3216	0.1258	0.3151
14	0.271	0.0342	0.0311	0.0623	-0.0142	-0.3161
15	-0.0375	0.0893	0.638	0.0392	-0.0665	0.0684
16	0.3355	0.0308	0.0004	0.0017	-0.3234	-0.1521
17	-0.005	-0.0052	0.0169	-0.0091	-0.0056	-0.0166
18	-0.0419	0.0266	0.1806	-0.0251	0.1186	-0.6417
19	0.1755	-0.0094	0.0475	-0.1115	0.1896	-0.2301
20	0.164	0.0085	0.0431	-0.2457	0.2672	-0.0815
21	0.3351	-0.0139	0.0399	-0.1526	0.1895	0.1308
22	0.3187	-0.0107	0.0392	-0.0863	0.2529	0.2365
23	0.3738	-0.0093	0.0421	-0.2474	0.2219	0.172
24	0.343	0.0271	-0.0002	-0.1493	-0.206	0.1874
25	0.043	0.0037	0.025	0.0064	-0.0495	0.024

Generally the name selected is governed by the largest correlations with the factor consistent with the nature of other variables having low correlation. With this procedure the following variables are recognizable to be associated with each of the factors as shown in Table 2.0. They are summarized below:

Factor I: variables: 16
(OCCU_CODE), 21 (S1_ DISTANCE), 22(S1_MODE), 23(S1_TYPE_PA), 24(S1_TRA_TI)

Factor II: variables:
1 (MALES), 2(FEMALES), 3(TOTAL_NO), 4(NO_MEM_STU)

Factor III: variables: 8
(TOTAL_VEHI), 11(RESIDENCE), 15(INCOME)

Factor IV: variables: 7
(VO_CAR), 9 (MONTH_INCO), 10(MONTH_TRAN)

Factor V: variables:
5(VO_CY), 6(TWO_WHLR), 12(AGE)

Factor VI: variables: 13
(SEX), 14(EDUCATION), 18(RAIL_BUS_P)

Factor I: Variables associated with this factor are: Access to public transportation, and other transport modes available for travel, availability of parking spaces, time taken to travel. All of them indicate opportunities available for the traveler that will enable to travel conveniently, hence this factor can be termed as "Mobility Opportunity Factor".

Factor II: Variables associated with this factor are: Number of male travelers, number of female travelers, and total numbers of members. They indicate the House hold family structure and hence can be termed as "Household Structure Factor".

Factor III: Variables associated with this factor are: The number of vehicles owned, Type of house living, Income of the household. These are related to the prosperity of the household and hence can be termed as "Prosperity factor".

Factor IV: Variables associated with this factor are: Number of cars, monthly Income and expenditure on transport of the traveler and hence can be termed as "Car ownership factor".

Factor V: is associated with Bicycles, Motor cycles and Age of the user. They indicate youth factor and can be termed as "Youth Factor".

Factor VI: is associated with Sex, Education, and Public Transport monthly pass system. Hence this factor can be termed as "Public transport Regular usage factor".

Since the main purpose of the Factor analysis is to understand what actually we are trying to measure, then the following criteria can best describe and applied to select most appropriate factor to measure for maximally capturing the socio economic status of the commuter for travel quantification purpose:

Variable 1: That should capture the "Availability of various transport facilities" to him for the travel

Variable II: That should explain his "Family size"

Variable III: That should explain his type of "Occupation status"

Variable IV: That should capture the type of "Own vehicle" he has for travel

Variable V: That should describe the "Age of the traveler"

Variable VI: That would explain the availability of cheaper "Public transportation"

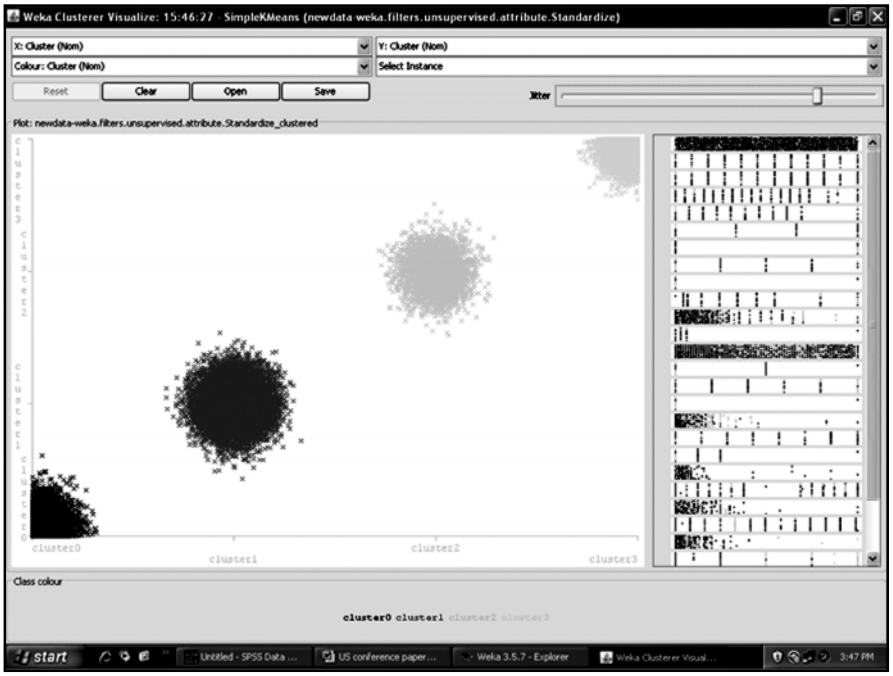
In other words the six basic traits covered in these factors are assumed to cover most of the information contained in several questions that are put to the traveler. In other words, the home Interview Survey Format can be revisited to reduce or restructure the number of questions so as to reduce the respondents' fatigue, and time for administering the questionnaire.

5.1 Inferences Drawn From Clustering

The data mining stage of KDD process involves applying various algorithms on the data to derive useful knowledge. It has to be remembered that no algorithm gives perfect results. Only the best and meaningful rules and clusters have been considered.

The data that is made available is related to Transportation Planning Study. The questionnaire has been designed and administered to understand the reasons that make the household members to choose different modes of transport. It is also intended to understand, as to why some use public transport when they have their own vehicle. The aim of the present study is to identify homogenous socio economic groups of people with similar trip making characteristics. This work attempts to mine the factors that can assist in understanding them and in turn to predict likely future scenarios, if these factors change over a period of time.

The Simple K-Means clustering algorithm has been applied to the full data, with 2, 4, 6 and 8 clusters as starting point. The resulting statistics has been compared. The four cluster formation has produced the lowest "within cluster sum of squared error" statistic (Figure 2). It is therefore decided to further examine these four clusters and the information contained in them. The following properties were associated with each cluster:



Weka Clusterer Visualize: 15:46:27 - SimpleKMeans (newdata-weka.filters.unsupervised.attribute.Standardize)

X: Cluster (Nom)

Y: Cluster (Nom)

Colour: Cluster (Nom)

Select Instance

Plot: newdata-weka.filters.unsupervised.attribute.Standardize_clustered

Figure 2: Cluster Output for k=4

First Cluster: Persons having high Income, having car ownership, persons having fewer dependents, or more earners, houses located away from public transport systems etc. In other words they belong to Prosperous families. One can identify the prosperity by associating Car with the house. Hence they can be categorised into "Car Owning Group".

Second Cluster: The second class of people are those whose incomes are slightly lower than the above class, but with more dependents, or less earners, with slightly less educational standards. They possess at least a Two Wheeler like, scooter, motor cycle, moped etc. This group can be considered as upper Middle Income group class and can be considered as those who own two wheelers. They can be categorised as "Motor cycle owning group".

Third Cluster: The third category of people mostly does not have a vehicle but, may have occasionally Two Wheeler, but they prefer to travel by public transport.

Their family size is slightly bigger, and this group can be considered as Lower Middle Income group people. May be considered has "No vehicle Owning group".

Fourth Cluster: The fourth category have low educational level, work in private sector, or work in some activity on daily wage basis. Mostly they have a bicycle if they work in fixed time schedule activity, or on contract basis. They have slightly lower type of residences. This group of people can be considered a having "Low Income or Bicycle owning group".

Accordingly these four classes can now be identified by their vehicle ownership levels, those who belong to "Vehicle Owning" and those who belong to "No Vehicle Owning" classes. Among the Vehicle owning group they can be further classified into those having a "Car ownership", "Two Wheeler Ownership", "Bicycle Owner ship" etc. In other words, the behaviour of these classes of people especially in the context of travel behaviour can be identified on that nomenclature. They are distinctly different from each other.

6. EVOLUTIONARY COMPUTATIONAL TECHNIQUES

From the above results it is evident that for complex problems like transportation, Data mining techniques are very useful to extract needful information. On the other hand if we use suitable advanced techniques , one can extract efficient information from these kind of multi dimensional datasets. Evolutionary computational techniques can answer the above.

Evolutionary computation techniques (ECTs) are random search methods simulating natural selection and evolution in the biological world. ECTs maintain a population of potential (or candidate) solutions to a problem and not just one solution [10]. There are five major ECTs:

- Genetic algorithm (GA)
- Genetic Programming (GP)
- Evolutionary Programming (EP)
- Particle Swarm Optimization(PSO)
- Differential evolution (DE)

These ECTs are population based approaches and due to this they can avoid being trapped in local optimum and consequently can often find global optimal solution. Thus, ECTs can be referred as global optimization algorithms. ECTs have already been successfully applied to a wide variety of optimization problem, for example: pattern recognition, scheduling, image processing, etc.

From the above results it is evident that for complex problems like transportation, data mining techniques are very useful to extract needful information. On the other hand if we use suitable advanced techniques , one can extract efficient information from these kind of multi dimensional datasets. Evolutionary computational techniques can answer the above.

With this study it is evident that we can extract some more useful information by using ECT's like PSO and DE for this Transportation Data.

7. CONCLUSION

The role and importance of Data mining in previously unexplored fields like transportation is thus well evident based on our work. We have explored and analyzed the role of data mining in the field of transportation. For this study we have taken a real time data set related to transportation planning called as HIS data, applied data preprocessing and data clustering techniques on this data and could able to draw useful inferences. Advanced data mining techniques like evolutionary computation techniques and others will be of great help to the transportation engineers and will pave the way for further research in data mining.

REFERENCES

[1] S. A. M. Weiss and N. Indurkhya, "Predictive Data Mining: A Practical Guide", *Morgan Kaufmann Publishers*, 1998.

[2] J. Han and M. Kamber, "Data Mining : Concepts and Techniques", *2nd Edition*, Elsevier, 2006.

[3] Overview of the KDD process-http://www2.cs.uregina.ca/~hamilton/courses/831/notes/kdd/1_kdd.html.

[4] Pyle, Dorian, "Data Preparation for Data Mining", Morgan Kaufmann Publishers, 1999.

[5] A. Jain, M. Murty, and P. Flynn, "Data Clustering: A review", *ACM Comput. Surv.*, **31**, No. 3, pp. 264–323, 1999.

[6] T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, "An Efficient *k*-means Clustering Algorithm: Analysis and Implementation", *IEEE Trans. Pattern Analysis and Machine Intelligence*, **24** (2002), 881-892.

[7] Development of Hyderabad Multimodal Suburban Commuter Transportation System, Government of Andhra Pradesh, 2004.

[8] Khist. C. Jotin, "Transportation Engineering: An Introduction", Prentice Hall Inc, ISBN 0-13-929274-8.

[9] http://www.cs.waikato.ac.nz/ml/weka/

[10] A.E. Eiben and J.E. Smith, "Introduction to Evolutionary Computing", Springer, 2003.