Algorithms Research 2013, 2(1): 18-23
DOI: 10.5923/j.algorithms.20130201.03

Prediction of Effective Mobile Wireless Network Data Profiling Using Data Mining Approaches

Ch. R. Phani Kumar[1,*], B. Uday Kumar[2], V. Malleswara Rao[2], Dsvgk Kaladhar[3]

[1]Department of ECE, GITAM University, Visakhapatnam, 530045, India
[2]Department of EIE, GITAM University, Visakhapatnam, 530045, India
[3]Department of Bioinformatics, GITAM University, Visakhapatnam, 530045, India

Abstract Mobile network analysis has a huge potential that provide insight into the relational dynamics of individuals. Machine learning and data mining techniques provide the behavior patterns of the mobile network data. The data transfer during all the days has produced good results in transfer of data starting from Day 1 to Day 22. Hierarchical clustering of the datasets for all the 1634 data examples in the mobile traffic dataset. Complete linkage dendrogram has been produced between 0 and 4.64. Two clusters have been produced from the present wireless mobile traffic datasets.

Keywords Mobile Wireless Network, Data Profile, Dataset, Data Mining

1. Introduction

Decision tree and neural network techniques can deliver accurate prediction models by using customer demographics, billing information, contract/service status, call detail records, and service change log[1]. Mobility prediction is one of the most essential issues that need to be explored for mobility management that maybe evaluated through simulation as compared with prediction methods in mobile computing systems[2].

Data collected from mobile phones have the potential to provide insight into the relational dynamics of individuals[3]. Advanced data mining techniques and a neural network algorithm can be combined successfully to obtain a high fraud coverage combined with a low false alarm rate[4]. Mobile mining is about finding useful knowledge from the raw data produced by mobile users consists of a set of static device and mobile device. Previous works in mobile data mining include finding frequency pattern and group pattern that builds a user profile based on past mobile visiting data, filters and to mine association rules[5].

Capacity in mobile networks will have to be assigned dynamically; necessitating development of new data mining techniques for understanding and predicting network load[6]. Active networks are a novel approach to network architecture in which the switches (or routers) of the network perform customized computations on the messages flowing through them. Active network perform user-driven computation at nodes within the network today, and the emergence of mobile code technologies that make dynamic network service innovation attainable[7].

The simplest approach to the analysis of sensor network data makes use of a centralized architecture where a central server maintains a database of readings from all the sensors from thousands of nodes, to manage large amount of data efficiently through data mining techniques[8]. The mobile user's behavior patterns, in which the location and the service are inherently coexistent, become more complex than those of the traditional web systems[9].

Network Value Analysis (NVA) as a way to analyse competitive ecosystems illustrates its application, the provision of mobile services and content can be explored to identify potential strategic implications for mobile operators[10].

2. Methodology

The Knowledge Discovery in Databases (KDD) methodology seems to be attractive on the analyze of large mobile databases. Following are the algorithms used to analyze the mobile traffic data containing 23 attributes and 1634 datasets.

2.1. Majority

Accuracy of classifiers is often compared to the "default accuracy", that is, the accuracy of a classifier which classifies all instances to the majority class. To fit into the standard schema, even this algorithm is provided in form of the usual learner-classifier pair. Learning is done by :obj:`MajorityLearner` and the classifier it constructs is an instance of :obj:`ConstantClassifier`.

.. class:: MajorityLearner

* Corresponding author:
chphanikr@gmail.com (Ch. R. Phani Kumar)
Published online at http://journal.sapub.org/algorithms

MajorityLearner will most often be used as is, without setting any parameters. Nevertheless, it has two.

Step1: .. attribute:: estimator_constructor

An estimator constructor that can be used for estimation of class probabilities. If left None, probability of each class is estimated as the relative frequency of instances belonging to this class.

Step2: .. attribute:: apriori_distribution

Apriori class distribution that is passed to estimatorconstructor if one is given.

Step3: .. class:: ConstantClassifier

ConstantClassifier always classifies to the same class and reports the same class probabilities.

Its constructor can be called without arguments, with a variable (for: obj:`class_var`), value (for :obj:`default_val `) or both. If the value is given and is of type: obj: `Orange.data.Value` (alternatives are an integer index of a discrete value or a continuous value), its attribute:obj:`Oran ge.data.Value.variable` will either be used for initializing :o bj:`class_var` if variable is not given as an argument, or checked against the variable argument, if it is given.

Step4: .. attribute:: default_val

Value that is returned by the classifier.

Step5: .. attribute:: default_distribution

Class probabilities returned by the classifier.

Step6: .. attribute:: class_var

Class variable that the classifier predicts.

2.2. Support Vector Machine (SVM)

Support vector machines (SVM) is a popular machine learning method with variants for classification, regression and distribution estimation that can learn a problem in a higher dimensional space through the use of a kernel trick. Integrated with Orange is a popular implementation by Chang and Lin, libsvm.[26]the fact from the multidimensional analytical geometry, namely that the distance between a hyperplane H = {x: wT x + w0 = 0) and a point y ∈ Rk is given by the formula d(H,y).

2.3. k-Nearest Neighbor

Algorithm

Step1:**findNearest**

A component that finds nearest neighbours of a given example. **K** Number of neighbours. If set to 0 (which is also the default value), the square root of the number of examples is used. **Changed:** the default used to be 1.

Step2:**RankWeight**

Enables weighting by ranks (default: true).

weighID

ID of meta attribute with weights of examples

Step3:**nExamples**

The number of learning examples. It is used to compute the number of neighbours if k is zero.

2.4. Random Forest

Algorithm

Step1: Let „N" be the number of training cases, and let „M" be the number of variables in the classifier. Choose mas input variables, to be used to determine the decision at a node of the tree; m should be much less than M.

Step2: Recurse a training set for this tree by choosing N times with replacement from all N available training cases (take a bootstrap sample). Rest of the cases to be estimated as error of the tree, by predicting their classes.

Step3: For each node in the tree, randomly choose m variables, which should be based on the decision at that node.

Step4: Calculate the best split based on these m variables in the training set. The value of m remains to be constant during forest growing. Random forest is sensitive to the value of m.

Step5: Each tree is grown to the largest extent possible, into many classification trees without pruning, in constructing a normal tree classifier.

Unsupervised algorithms such as hierarchial clustering and SOM provide the problem of trying to find hidden structure in unlabeled data. Learning useful structure without labeled classes, optimization criterion, feedback signal, or any other information beyond the raw data can be processed in this methods

2.5. Hierarchical Clustering

Algorithm

Step 1: Cluster data with high S_{min}→ 1st hierarchical level

Step 2: Decrease S_{min} (stop at S_{min}=0)

Step 3: Treat cluster centroids as data tuples and recluster, creating next level of hierarchy, then repeat steps 2 and 3.

2.6. SOM (Self organizing maps)

Algorithm

Step1: Randomize the map's nodes' weight vectors

Step 2: Grab an input vector $\mathbf{D}(\mathbf{t})$

Step 3: Traverse each node in the map

a. Use the Euclidean distance formula to find the similarity between the input vector and the map's node's weight vector

b. Track the node that produces the smallest distance (this node is the best matching unit, BMU)

Step 4: Update the nodes in the neighborhood of the BMU (including the BMU itself) by pulling them closer to the input vector

$\mathbf{Wv}(s + 1) = \mathbf{Wv}(s) + \Theta(u, v, s)\, \alpha(s)(\mathbf{D}(t) - \mathbf{Wv}(s))$

Step 5: Increase s and repeat from step 2 while $s < \lambda$

Sieve graphs provide mean, sorting, skewness, and kurtosis related to the mobile traffic data.

3. Results

The mobile traffic datasets have been analysed using data mining techniques. The accuracy and errors in the examples were found by using various classification algorithms.

The datasets were also visualized using plotting techniques like the distribution plot, sieve graph, etc.

Figure 1 shows that the traffic levels had 96% utilization during data transfer in mobile technology predicted in this dataset.

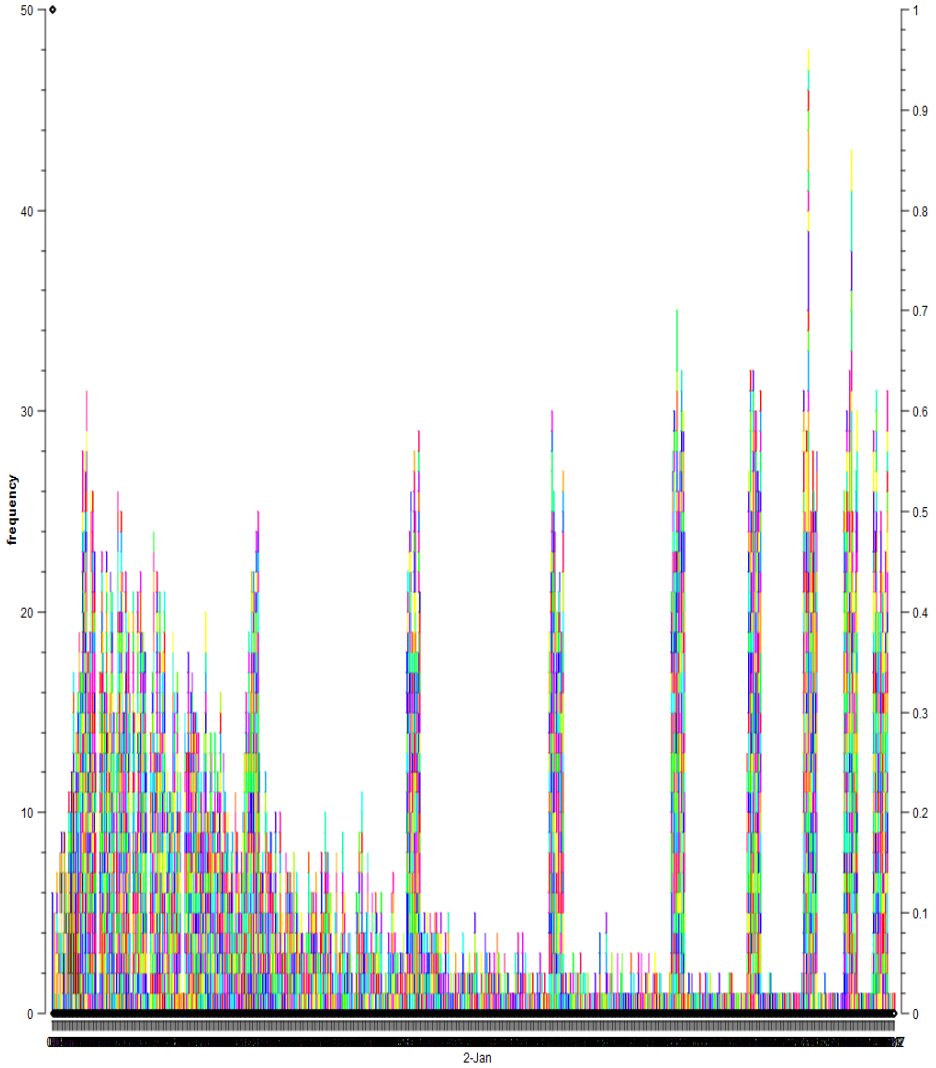

Figure 1. Distribution plot from mobile traffic datasets

Table1 shows the classification of mobile traffic datasets. As the RMSE is less for kNN, the accuracy of the datasets for kNN was maximum. As R2 is zero for majority, so this algorithm has produced less predicted accuracy for our present mobile traffic datasets.

Table 1. Classification of mobile traffic dataset

	MSE	RMSE	MAE	RSE	RRSE	RAE	R2
Majority	186.3	13.6	9.09	1.0	1.0	1.0	0
SVM	9.06	3.01	1.44	0.04	0.22	0.15	0.95
KNN	4.52	2.13	1.34	0.02	0.15	0.15	0.97
Random Forest	6.37	2.52	1.59	0.03	0.18	0.17	0.96

Figure 2 shows the SVM patterns for mobile traffic datasets. This unsupervised learning is very useful when compared with supervised learning such as classification algorithms as there are more number of datasets in high skewness (ranges from 49.83 – 599.67). There is less number of utilization observed by using this SOM method.

Figure 3 shows the visualization for the mobile traffic datasets. The data transfer during all the days has produced good results in transfer of data starting from Day 1 to Day 22

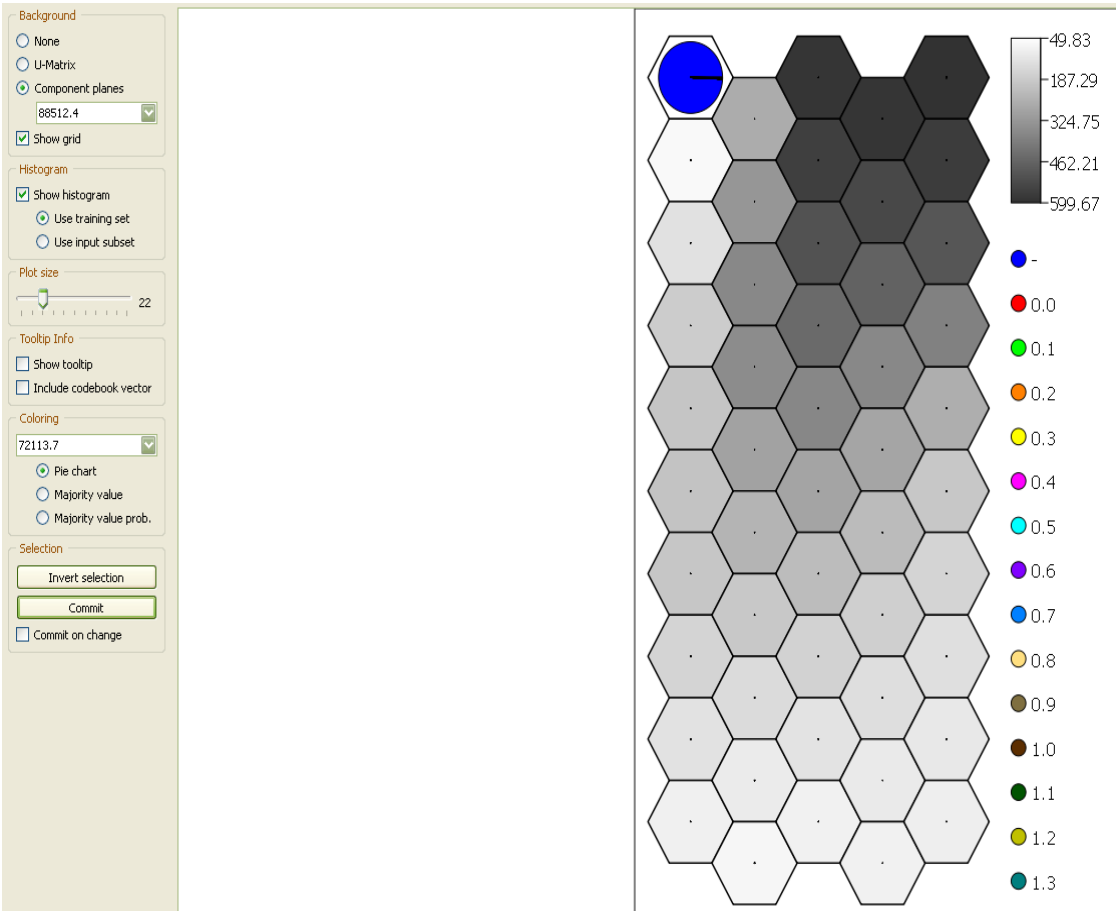

Figure 2. SOM

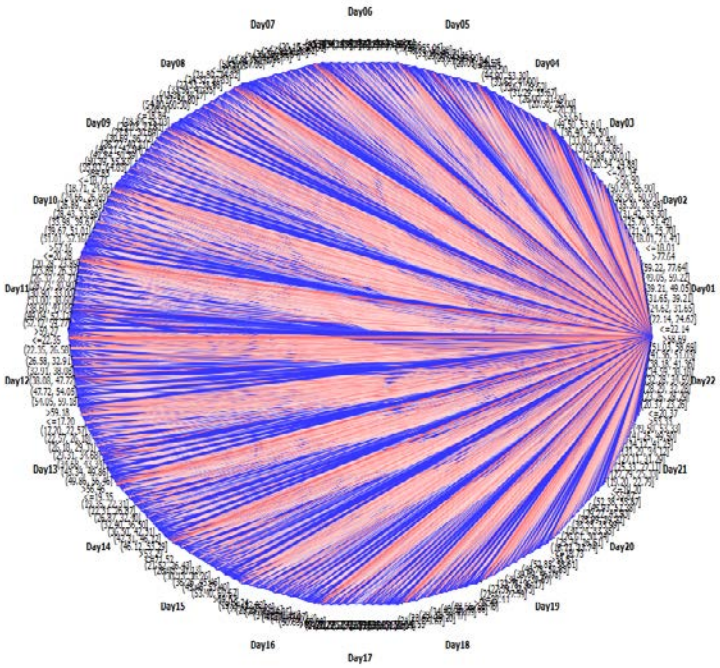

Figure 3. Seive graph

Figure 4 shows the hierarchical clustering of the datasets for all the 1634 data examples in the mobile traffic dataset. Complete linkage dendrogram has been produced between 0 to 4.64. Two clusters has been produced from the present wireless mobile traffic datasets.

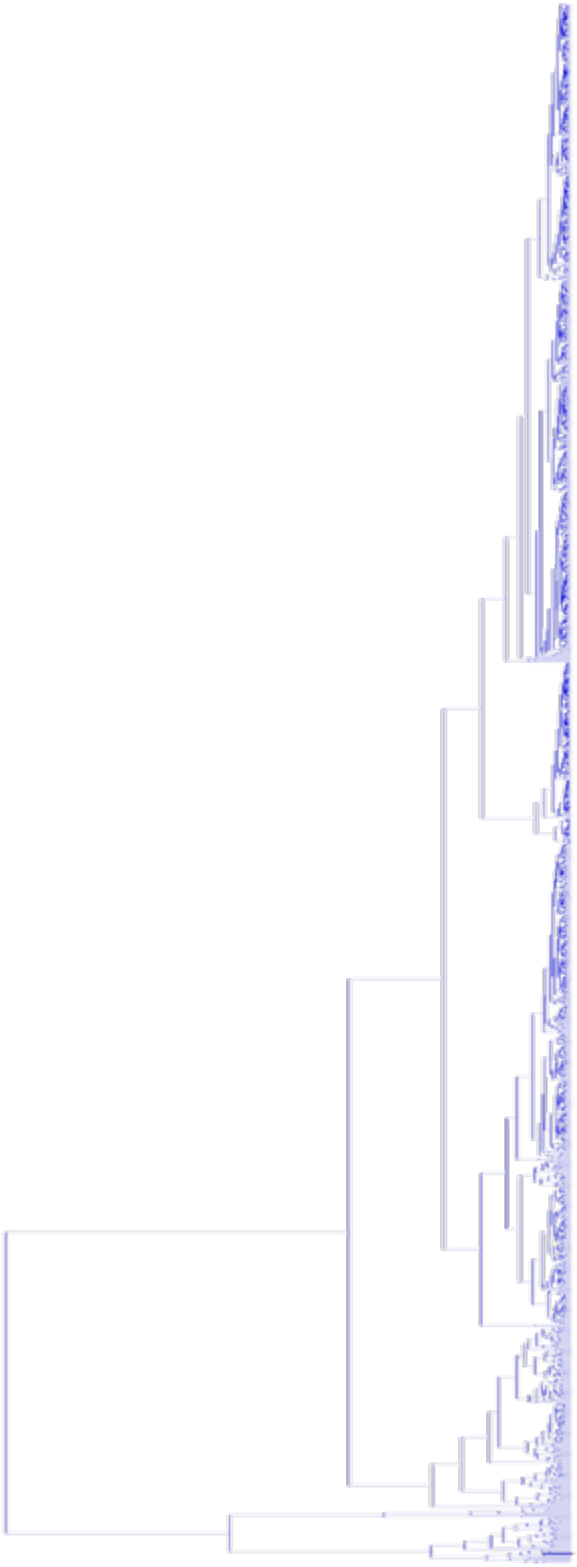

Figure 4. Hierarchical clustering

4. Discussion

Data mining as we use the term is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules[11]. Data mining will be able to make a high impact in the area of integrated data fusion and mining in ecological/environmental applications, especially when involving distributed/decentralized data sources like autonomous mobile sensor networks[12]. Knowledge discovery in databases is a rapidly growing field, whose development is driven by strong research interests as well as urgent practical, social, and economical needs[13].

The development of wireless and web technologies has allowed the mobile users to request various kinds of services by mobile devices at anytime and anywhere. Helping the users obtain needed information effectively is an important issue in the mobile web systems[14]. In this correspondence, we address the issue of efficiently mining multilevel and location-aware associated service patterns in a mobile web environment. In terms of multilevel concept, we consider the complex problem that locations and services are of hierarchical structures[15].

Mobility prediction is one of the most essential issues that need to be explored for mobility management in mobile computing systems[16].As deregulation, new technologies, and new competitors open up the mobile telecommunications industry, churn prediction and management has become of great concern to mobile service providers[17].

Because the radio spectrum is limited, managing the limited amount of resources is an important issue, especially for high-speed data applications thus making data mining prediction models important[18]. New data mining techniques involve mining calling path patterns in global system for mobile communication (GSM) networks[19].

A neural network algorithm called the self-organizing map, together with a conventional clustering method like the k-means, can effectively be used to simplify and focus network analysis. It is shown that these algorithms help in visualizing and grouping similarly behaving cells. Thus, it is easier for a human expert to discern different states of the network. This makes it possible to perform faster and more efficient troubleshooting and optimization of the parameters of the cells[20].

5. Conclusions

Thus we have predicted the accuracies and errors in mobile traffic data using various classification algoritms. From the study, we come to the conclusion that the Majority algorithm is the one with the least accuracy and the k-Nearest Neighbors algorithm provides the maximum accuracy,i.e., least error. Thus we propose to use the k-Nearest Neighbours alogorithm or also SVM and Random forest algorithm can be used to make predictions in mobile traffic usage which provide an acceptable level of accuracy.

ACKNOWLEDGEMENTS

Authors would like to thank management and staff of

GITAM University, India for their kind support in bringing out the above literature and providing lab facilities.

REFERENCES

[1] Shin Yuan Hunga, David C. Yenb, Hsiu-Yu Wangc, "Applying data mining to telecom churn management". Expert Systems with Applications, vol. 31, no. 3, pp. 515–524, 2006.

[2] Gökhan Yavaşa Dimitrios Katsarosb, Özgür Ulusoya,Yannis Manolopoulosb, "A data mining approach for location prediction in mobile environments". Data & Knowledge Engineering, vol. 54, no. 2, pp. 121–146, 2005.

[3] Nathan Eagle, Alex (Sandy) Pentland, David Lazer,"Inferring friendship network structure by using mobile phone data". PNAS vol. 106, no. 36, pp. 15274-15278, 2009.

[4] R. Brause, T. Langsdorf, M. Hepp,"Neural data mining for credit card fraud detection". Tools with Artificial Intelligence, 11th IEEE International Conference, pp. 103 – 106, 1999.

[5] Jen Ye Goh, David Taniar, "Mobile Data Mining by Location Dependencies". Intelligent Data Engineering and Automated Learning – IDEAL Lecture Notes in Computer Science,vol. 3177, pp. 225-231, 2004.

[6] J. Han, R. B. Altman, V. Kumar, H. Mannila, D. Pregibon, "Emerging scientific applications in data mining". Communications of the ACM, vol. 45, pp. 54-58, 2002.

[7] D. L. Tennenhouse, J. M. Smith, W. D. Sincoskie, D. J. Wetherall, G. J. Minden, "A survey of active network research. Communications Magazine". IEEE, vol. 35, pp. 80-86, 1997.

[8] G. Bontempi, Y. Le Borgne, "An adaptive modular approach to the mining of sensor network data". In Proceedings of 1st International Workshop on Data Mining in Sensor Networks as part of the SIAM International Conference on Data Mining, pp. 3-9, 2005.

[9] V. S. Tseng, K. W. Lin, "Efficient mining and prediction of user behavior patterns in mobile web systems". *Information and software technology*, vol. *48*, pp. 357-369, 2006.

[10] J. Peppard, A. Rylander, "From Value Chain to Value Network:: Insights for Mobile Operators". *European Management Journal*, vol. *24*, pp. 128-141, 2006.

[11] M. J. Berry, G. S Linoff, "Data mining techniques: for marketing, sales, and customer relationship management". Wiley Computer Publishing, 2004.

[12] Q. Yang, X. Wu, "10 challenging problems in data mining research". International Journal of Information Technology & Decision Making, vol. 5, pp. 597-604, 2006.

[13] M. Goebel, L. Gruenwald, "A survey of data mining and knowledge discovery software tools". *ACM SIGKDD Explorations Newsletter*, vol. *1*1, pp. 20-33, 1999.

[14] V. S. Tseng, K. W. Lin, "Efficient mining and prediction of user behavior patterns in mobile web systems". *Information and software technology*, vol. *48*, pp. 357-369, 2006.

[15] S. M. Tseng, C. F. Tsui, "Mining multilevel and location-aware service patterns in mobile web environments". *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. *34*, pp. 2480-2485, 2004.

[16] G. Yavas, D. Katsaros, Ö. Ulusoy, Y. Manolopoulos, "A data mining approach for location prediction in mobile environments". *Data & Knowledge Engineering*, vol. *54*, pp. 121-146, 2005.

[17] C. P. Wei, I. Chiu, "Turning telecommunications call details to churn prediction: a data mining approach". *Expert systems with applications*, vol. *23*, pp. 103-112, 2002.

[18] J. L. Chen, "Resource allocation for cellular data services using multiagent schemes". Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 31, pp. 864-869, 2001.

[19] A. J. Lee, Y. T. Wang,"Efficient data mining for calling path patterns in GSM networks". Information Systems, vol. 28,pp. 929-948, 2003.

[20] J. Laiho, K. Raivio, P. Lehtimaki, K. Hatonen, O. Simula, "Advanced analysis methods for 3G cellular networks". *Wireless Communications, IEEE Transactions on*, vol. *4*, pp. 930-942, 2005.