Implementation of Data Mining Decision Tree Algorithms on Mobile Computing Environment

Neha Sobti, Ketki Arora

Abstract-The idea of complex activity for characterizing the continuously changing complex behavior patterns of mobile users. For the purpose of data management, a complex activity is modeled as a sequence of location movement, service requests, the co-occurrence of location and service, or the interleaving of all above. An activity may be composed of sub activities. We, therefore, propose new methods for complex activity mining, incremental maintenance, online detection and proactive data management based on user activities. In particular, we devise prefetching and pushing techniques with cost-sensitive control to facilitate predictive data allocation. Preliminary implementation and simulation results demonstrate that the proposed framework and techniques can significantly increase local availability, conserve execution cost, reduce response time, and improve cache utilization. Different activities may exhibit dependencies that affect user behaviors. We argue that the complex activity concept provides a more precise, rich, and detail description of user behavioral patterns which are invaluable for data management in mobile environments. Proper exploration of user activities has the potential of providing much higher quality and personalized services to individual user at the right place on the right time. With the help of data mining algorithms, we will try to reduce execution time, find correctly classified instance, reduce error rate and improve accuracy.

Keywords: ID3, DTNA, Mobile environment, data mining algorithms

I. INTRODUCTION

A sequence of user behaviors that may be composed of location-only patterns, service-only patterns, location-service pairs, or the inter- leaving of all above, is called a complex activity. We argue that the activity concept provides a more precise, rich, and detail description of user behavioral patterns which are invaluable for data management in mobile environments. Proper exploration of such activities enables data management system to predict the user's next move, intended service or both for providing much higher quality and personalized services to individual user at the right place on the right time. Such kind of advanced information services call for new methods of complex activity mining, incremental maintenance, online detection, and proactive data management based on user activities. We propose new pattern mining and online processing algorithms to the discovery and maintenance of complex activities in mobile environments. Furthermore, we devise prefetching, pushing, and handoff techniques with cost-sensitive control to facilitate predictive data allocation.

Classification is an important task in data mining. Its purpose is to set up a classifier model and map all the samples to a certain class which can provide much convenience for people to analyze data further more.

Classification belongs to directed learning, and the main methods include decision tree, Bayesian classification, neural network, genetic algorithm and rough set etc. while the mostly used decision tree algorithms are ID3 and C4.5 presented by J. R. Quinlan. Many classical methods are improved based on them. But there still exist some problems like multivalue bios and the pruning of trees etc. which we need to research deeply. This paper improves the traditional algorithms like ID3, and presents a new synthesized algorithm DTNA for mining large-scale high dimensional datasets. The basic idea of DTNA is shown as follows: first introduce PCA to analyze the relevancy between features and replace the whole dataset with several countable composite features; then improve ID3 to part the set into several clusters which can be the pretreatment of other algorithms and achieve the reduction of sample scale. Decision Tree Method Decision tree is one of the important analysis methods in classification. It builds its optimal tree model by selecting important association features. While selection of test attribute and partition of sample sets are two crucial parts in building trees. Different decision tree methods will adopt different technologies to settle these problems. Traditional algorithms include ID3, C4.5, CART, SPRINT, SLIQ etc. ID3 is the representation of decision tree method. It is easy to understand and has fast classified speed which is applicable to large datasets. Many decision tree algorithms are improved based on it, like ID3. But these algorithms more or less have some problems in selection of test features, type of samples, memory utilization of data and the pruning of trees etc. Presently, researchers have present many improvements. Applied genetic algorithms to prune decision trees.

ID3 builds decision trees from a set of training data in the same way as C4.5 using the concept of information entropy[6]. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The ID3 algorithm then recurses on the smaller sub lists.

The "Iterative Dichotomiser Tree" (ID3) works only with categorical input and output data. ID3 grows trees splitting on all categories of an attribute, thus producing shallow and wide trees. It grows tree classifiers in three steps:

1. Splits creation in form of multiway splits, i.e. for every attribute a single split is created where attributes' categories are branches of the proposed split.

Manuscript Received on May 2014.

NehaSobti, Research Scholar, Lovely Professional University
KetkiArora, Assistant Professor, Lovely Professional University

2. Evaluation of best split for tree branching based on *information gain measure*, and
3. Checking of the stop criteria, and recursively applying the steps to new branches. These three steps are iterating and are executed in all nodes of the decision tree classifier. The information gain measure is based on the well-known Shannon entropy measure.

$$E(X,S) = \sum_{j=1}^{K} \left(\frac{|S_j|}{|S|} \cdot E(S_j) \right)$$

$$I(X,S) = E(S) - E(X,S)$$

Proposed Algorithm:

(1) create a node N;
(2) if samples are all of the same class, C then
(3) return N as a leaf node labeled with the class C;
(4) if attribute-list is empty then
(5) return N as a leaf node labeled with the most common class in samples;
(6) select test-attribute, the attribute among attribute-list with the highest information gain;
(7) label node N with test-attribute;
(8) for each known value a_i of test-attribute;
(9) grow a branch from node N for the condition test-attribute = a_i;
(10)lets$_i$ be the set of samples in samples for which test-attribute = ai; // a partition
(11)ifs$_i$ is empty then
(12)attach a leaf labeled with the most common class in samples;
(13)else attach the node returned by Generate_decision_tree (s_i, attribute-list- test-attribute);

The basic strategy is as follows:

The tree starts as a single node representing the training samples (step 1). If the samples are all of the same class, then the node becomes a leaf and is labeled with that class (steps 2 and 3). Otherwise, the algorithm uses an entropy-based measure known as information gain as a heuristic for selecting the attribute that will best separate the samples into individual classes (step 6). This attribute becomes the "test" or "decision" attribute at the node (step 7). (All of the attributes are categorical or discrete value. Continues-valued attribute must be discretized.) A branch is created for each known value of the test attribute, and the samples are partitioned accordingly (steps 8-10). The algorithm uses the same process recursively to form a decision tree for the samples at each partition. Once an attribute has occurred at a node, it need not be considered in any of the node's descendents (step 13). The recursive partitioning stops only when any one of the following conditions is true: o All the samples for a given node belong to the same class (steps 2 and 3), or o There are no remaining attributes on which the samples may be further partitioned (step 4). In this case, majority voting is employed (step 5). This involves converting the given node into a leaf and labeling it with the class in majority among samples. Alternatively, the class distribution of the node samples may be stored.

o There are no samples for the branch test-attribute = a_i (step 11). In this case, a leaf is created with the majority class in samples (step 12).

ToolsofDataCollection&Analysis

Weka, NetBeans, Java, Tanagra

WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a stateofthe-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data preprocessing, classification, regression, clustering and association rules; It also includes visualization tools. The new machine learning schemes can also be developed with this package. WEKA is an open source software issued under General Public License. The data file normally used by Weka is in ARFF file format, which consists of special tags to indicate different things in the data file (foremost: attribute names, attribute types, attribute values and the data). The main interface in Weka is the Explorer. It has a set of panels, each of which can be used to perform a certain task. Once a dataset has been loaded, one of the other panels in the Explorer can be used to perform further analysis.

Attributes of Data set:

1.	Strength
2.	Authentication type
3.	Threshold
4.	Frequency
5.	Antennas
6.	GPS Signals
7.	Satellites
8.	Transmitted Frame
9.	Multicast Frame
10	ACK failure count
11.	Services

Implementation and Result

Simulator implementation for collection of mobile environment data set:

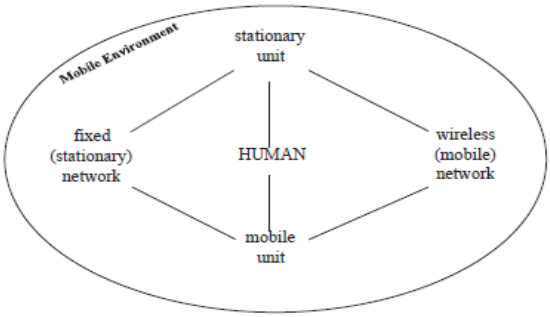

Figure 1: Data Set Collection based on Mobility

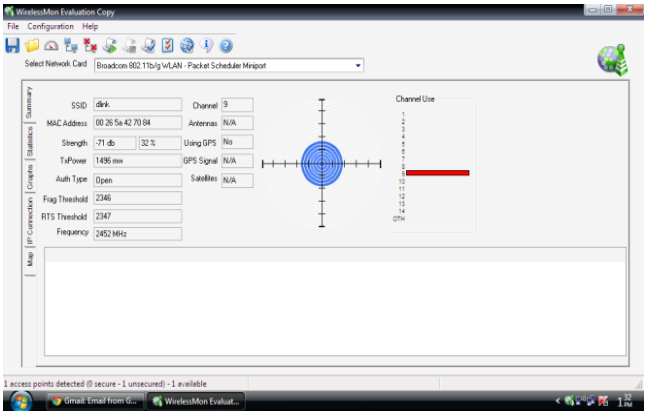

Figure 2: Strength, Auth. Type, Threshold and Frequency of the Mobile Network

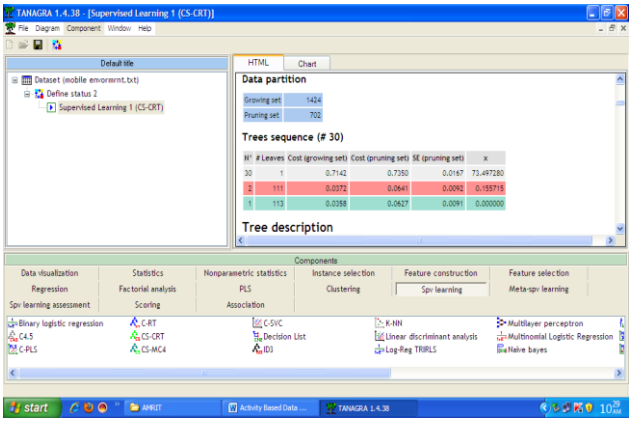

Figure 3: Calculate no of Nodes

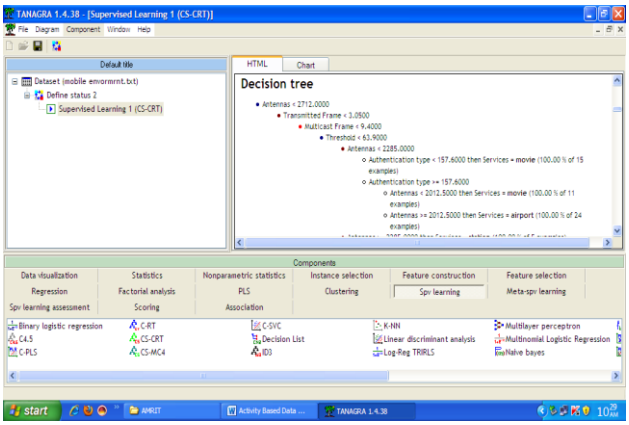

Figure 4: Create Decision Tree

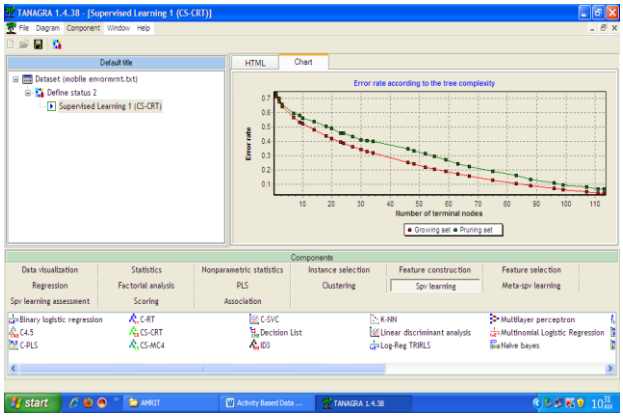

Figure 5: Show Growing Set and Pruning Set

Table: Analysis between Execution Time, Correctly Classified Instances, Error Rate

	ID3	DTNA
Execution Time	8.41 Seconds	0.19 Seconds
Correctly classified instances	54% (2526)	100% (4604)
Error rate	99%	0%

II. CONCLUSION

In this Research, I wanted to highlight the approaches for creating a decision tree. They are mainly available into academic tools from the machine learning community. I note that they are an alternative quite credible to decision trees and predictive association rules, both in terms of accuracy than in terms of processing time. After analysis Order ID3 and Enhanced algorithm is more suitable to find accurate and consuming less access time to mine data with minimum error rate. so enhanced algorithm is a best algorithm for mining a data on mobile environment data set.

REFERENCES

1. Xiang Lian, Student Member, IEEE, and Lei Chen, Member, IEEE, "Ranked Query Processing in Uncertain Databases", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO.3, MARCH 2010.
2. Stavroula G. Mougiakakou, Member, IEEE, "SMARTDIAB: A Communication and Information Technology Approach for the Intelligent Monitoring, Management and ollow-up of Type 1 Diabetes Patients", IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 14, NO. 3, MAY 2010.
3. Eric Hsueh-Chan Lu, Vincent S. Tseng, Member, IEEE, "Mining Cluster-Based Temporal Mobile Sequential Patterns in Location-Based Service Environments" , IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 6, JUNE 2011.
4. Mark N. Gasson, EleniKosta, Denis Royer, Martin Meints, and Kevin Warwick, "Normality Mining: Privacy Implications of Behavioral Profiles Drawn From GPS Enabled Mobile Phones", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 41, NO. 2, MARCH 2011.
5. Tzung-Shi Chen, Member, IEEE, Yen-Ssu Chou, and Tzung-Cheng Chen, "Mining User Movement Behavior Patterns in a Mobile Service Environment", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 42, NO. 1, JANUARY 2012.
6. R. Agrawal, T. Imielinski, and A. Swami, "Mining Associations between Sets of Items in Massive Databases," Proc. ACM SIGMOD, pp. 207-216, May 1993.
7. R. Agrawal and J. Shafer, "Parallel Mining of Association Rules," IEEE Trans. Knowledge and Data Eng, vol. 8, no. 6, pp. 866-883, Dec. 1996.
8. R. Agrawal and R. Srikant, "Mining Sequential Patterns," Proc. 11th Int'l Conf. Data Eng., pp. 3-14, Mar. 1995.
9. B. Bruegge and B. Bennington, "Applications of Mobile Computing and Communication," IEEE Personal Comm., pp. 64-71, Feb. 1996.
10. M.-S. Chen, J. Han, and P.S. Yu, "Data Mining: An Overview from Database Perspective," IEEE Trans. Knowledge and Data Eng., vol. 8, no 6, pp. 866-883, Dec. 1996..
11. M.-S. Chen, J.-S.Park, and P.S. Yu, "Efficient Data Mining for Path Traversal Patterns," IEEE Trans. Knowledge and Data Eng., vol. 10, no. 2, pp. 209-221, Apr. 1998.
12. D.W. Cheung, V.T. Ng, W. Fu, and Y. Fu, "Efficient Mining Association Rules in Distributed Databases," IEEE Trans. Knowledge and Data Eng., vol. 8, no. 6, pp. 911-922, Dec. 1996.
13. T.H. Cormen, C.E. Leiserson, and R.L. Rivest, Introduction to Algorithm. MIT Press, 1989.

14. SouptikDatta, Chris R. Giannella, and HillolKargupta, Senior Member, IEEE, "Approximate Distributed K-Means Clustering over a Peer-to-Peer Network" , IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 10, OCTOBER 2009.
15. Kwong-Sak Leung, Kin Hong Lee, Jin-Feng Wang, "Data Mining on DNA Sequences of Hepatitis B Virus", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 8, NO. 2, MARCH/APRIL
16. Shady Shehata, Member, IEEE, FakhriKarray, Senior Member, IEEE, and Mohamed S. Kamel, Fellow, IEEE, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 10, OCTOBER.
17. Lijun Wang, ManjeetRege, Ming Dong, Member, IEEE, and Yongsheng Ding, Senior Member, IEEE, "Low-Rank Kernel Matrix Factorization for Large-Scale Evolutionary Clustering" , IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 6, JUNE 2012
18. Eghbal G. Mansoori, "FRBC: A Fuzzy Rule-Based Clustering Algorithm", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 19, NO. 5, OCTOBER.
19. Ji Dan, QiuJianlin, "A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree", 2010 10th IEEE International Conference on Computer and Information Technology (CIT 2010).

AUTHORS PROFILE

Neha Sobti, has done Msc Compsc from hansraj mahila maha vidyalaya,Jalandhar in 2008.Currently,she is a Mtech student in department of computer science and engineering at Lovely professional university LPU,chaheru,India.

Ketki Arora, has done Mtech Compsc from Lala Lajpat Rai Institute of Engineering and technology LLRIET,Moga,India.Currently,she is an assistant professor at Lovely Professional UniversityLPU,chaheru,India .