An Overview on Mobile Data Mining

D. Natarajasivan
Assistant Professor
Department of Computer Science and Engineering
Annamalai University

M. Govindarajan, Ph.D
Assistant Professor
Department of Computer Science and Engineering
Annamalai University

ABSTRACT

In early days the mobile phones are considered to perform only telecommunication operation. This scenario of mobile phones as communication devices changed with the emergence of a new class of mobile devices called the smart phones. These smart phones in addition to being used as a communication device are capable of doing things that a computer does. In recent times the smart phone are becoming more and more powerful in both computing and storage aspects. The data generated by the smart phone provide a means to get new knowledge about various aspects like usage, movement of the user etc. This paper provides an introduction to Mobile Data Mining and its types.

General Terms

Data Mining, Classification, Clustering

Keywords

Mobile Data Mining, Location Based Service, Behavior Analysis, Movement Prediction, Intrusion Detection.

1. INTRODUCTION

Data mining is a novel approach which is applied to large data to get useful and previously unknown information. This information is generally regarded as the knowledge that provide people with ability to understand the nature of the data and take decisions based on the acquired knowledge. Data mining is applied in the places where there is large data but less information available for the user.

Mobile phones apart from being a communication device perform many tasks such as internet browsing, game playing, multimedia functionalities and application usage. This leads the mobile phones to collect large amount of data regarding their usage from which new knowledge can be obtained by mining. Mobile data mining is an approach which analyzes the data collected from the mobile phones and provides useful knowledge. In mobile data mining there are various aspects in which the data can be mined. In this paper some of the major areas of the mobile data mining are analyzed.

Section 2 presents a brief overview on various types of Mobile Data Mining. Section 3 describes the data and the techniques used for mobile data mining. Section 4 concludes the study.

2. MOBILE DATA MINING

Mobile data mining is a data mining approach that uses data collected from the mobile phones. The data collected by the mobile phones varies from simple call log to more application specific usage data. Figure 1 describes the various data that are collected by the mobile phone.

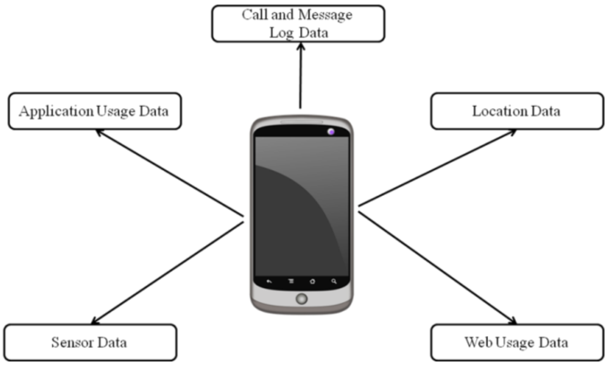

Fig 1: Data Generated in Mobile Phone

Using the data that are available on the mobile phones the data mining activity can be categorized as Location Prediction and Location Based Service, Behavior Analysis and Intrusion Detection.

2.1 Location Prediction and Location Based Service

Location prediction [2] provides a means of identifying the next location that a mobile user will be visiting based on the movement history of the mobile user. When the user's next location is predicted in advance it will lead to effective resource allocation with regards to the mobile operators. Location prediction also helps in environment where location based queries tends to come from the user of the mobile phone. If the next location is predicted the location based queries can be effectively satisfied.

Three main steps are taken in predicting the location of a mobile phone. In the first step the history of the mobile users movements are obtained and are organized in an effective manner. In the second step patterns are generated using an appropriate method from the mobile movement data. In the third step real time location information of a user is provided and the next location the user will be visiting is obtained based on the pattern generated in the second step.

Location Based Service (LBS) [5] are application specific services that are rendered by an application on the user's mobile phone based on the current location in which the user is active. LBS increase the revenue of an application vendor by customizing the application based on the usage of that particular application in a specific location. Some of the major LBS provided by a mobile application are Traffic Information, Restaurant Information, nearest Hospital and nearest Bank. Figure 2 depicts the general architecture of the Service request and LBS Reply.

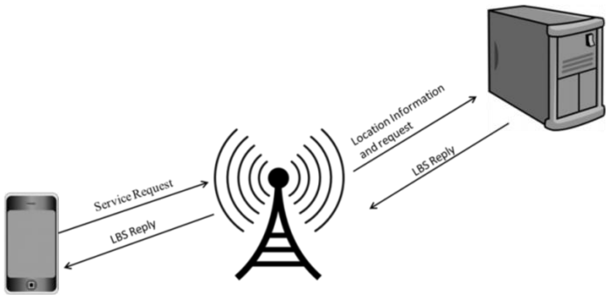

Fig 2: LBS Application Request and Reply

Using data mining techniques the application that provide LBS promote the appropriate information that a user needs at a particular location in advance. To provide effective LBS, the application vendor has to collect all the data regarding the application user such as the location of the user, information requested by the user etc. Using an appropriate data mining method the request patterns are identified. When there is a request by the application user the application server sends the appropriate LBS based on the pattern identified.

2.2 Behavior Analysis

Behavior analysis [6], analyze the behavior of a mobile user by considering various activities carried out by the user on the mobile phone and its social impact. Behavior analysis makes the mobile phone tailored for a person based on the interaction made by the person. Behavior of a mobile user may be the social interactions in a community or it may be the personality of the user or it may be an organizational behavior.

In behavior analysis the user interaction with the mobile phones gives an overview of what a user is doing at a particular context, such as waiting in the hospital, doing shopping or waiting for train. The context-awareness of user interactions reflects the habits of mobile users. Therefore, it is implied that the associations between user interaction records and the corresponding contexts can be used to characterize user habits. For example, the following associations can characterize the habits of a person:

- Given the context that a *person is waiting for a train in the evening of a work day and the surrounding environment is noisy*, the person usually *plays games*.
- ·Given the context that *a person is going to work in train*, the person usually *listens to a particular FM Radio Station*.
- ·Given the context that a *person is visiting a tourist attraction*, the person usually *shoots pictures*.

This kind of association is called as *behavior patterns*. A wide range of context-aware services, such as context-aware recommendation, context-aware User Interface adaption, can be improved by understanding the habits of mobile users from their behavior patterns.

2.3 Intrusion Detection

Intrusion detection [8] helps the mobile phone user to identify any malicious application that resides in the mobile phone or an attack by virus or worm. As the mobile phones are becoming more efficient in terms of speed and space, the users tend to store sensitive information in the mobile phones. This needs security to be imposed on the phone in order to maintain confidentiality. Traditional system based security mechanisms will not provide adequate security as they are hard to implement owing to the resource constraint of the mobile phone. There are two types of threats that are detected using the intrusion detection mechanism namely anomaly detection and misuse detection.

Main aim of anomaly detection [4] is to identify the situations that are unusual. It is an important tool for detecting rare events that are significant but hard to find, like fraud, network intrusion, etc. It can detect previously unknown intrusions as it does not require prior knowledge of intrusion. The main drawback of this approach is that it may not be able to describe what the attack is and may have high false positive rate.

Misuse detection systems use patterns that are gathered from previous known attacks or weak spots of the system to compare and identify know intrusions. In misuse detection, the information gathered is analyzed by the Intrusion Detection System (IDS) to identify matching attack signatures from the large database of attack signature. In essence, the IDS look for a specific attack signature that has already been identified as intrusion. For example, signatures for attacks that try to identify the password of a user can be “if there are more than 3 failed login attempts within 3 minutes”. This approach can detect instances of know attacks accurately and efficiently, but it lacks the ability to detect attack that are not previously known.

Data mining based Intrusion detection system on mobile phones are implemented in a resource effective manner as the detection system collects the needed data from the mobile phone and send the data to the server. The server analyzes the data and informs the detection system if there is any abnormality. This ultimately reduces the processing overhead in the mobile phone.

3. MOBILE DATA MINING TECHNIQUES

3.1 Location Prediction and Location Based Service

Location Prediction and Location Based Services needs to use the spatio-temporal data collected from the mobile phones. The spatio-temporal data contains the information of location and time as a pair. The spatio-temporal data is hard to mine as the search space for finding the knowledge is very high. The basic approach for mining this kind of data is to identify the frequent patterns in the data.

A Frequent Pattern [12] is defined over a database of sequences D, where each element of each sequence is a time-stamped set of items i.e., an *itemset*. Time-stamps determine the order of elements in the sequence. Then, the Frequent Pattern problem consists in finding all the sequences that are frequent in D, i.e., appear as subsequence of a large percentage of sequences of D. A sequence $\alpha = \alpha_1 \rightarrow \ldots \rightarrow\alpha_k$ is a subsequence of $\beta = \beta_1 \rightarrow \ldots\rightarrow\beta_m$ $(\alpha\leq\beta)$ if there exist integers $1 \leq i_1< . . . <i_k\leq m$ such that $\forall_{1\leq n\leq k}\ \alpha_n\subseteq\beta_{in}$. Then the support $supp_D(S)$ of a sequence S can be defined as the percentage of transactions $T \in D$ such that $S \leq T$, and say that S is frequent w.r.t. threshold s_{min} if $supp_D(S) \geq s_{min.}$

The spatio-temporal sequence used in location prediction is a sequence of triples

$$T = <x_0, y_0, t_0>, . . . ,<x_n, y_n, t_n> \qquad (1)$$

where t_i (i = 0. . . n) denotes a time-stamp such that $\forall_{0<i<n}t_i< t_{i+1}$ and (x_i, y_i) are points in $\mathbf{R}^2$. Intuitively, each triple $< x_i, y_i, t_i>$ indicates that the object is in the position (x_i, y_i) at time t_i. The Sequence Pattern Mining is used to identify the patterns in the spatio-temporal sequence dataset. This pattern is used in prediction the location of the user.

The Location Based Service uses mobile transaction sequence for predicting the service that the user is going to use. A mobile transaction sequence of a user with length equal to m is a sequence as given below

$$S = <(t_1, l_1, s_1), (t_2, l_2, s_2), \dots, (t_m, l_m, s_m)> \qquad (2)$$

where item (t_i, l_i, s_i) represents the user requests service s_i in location l_i at time t_i and $t_i < t_{i+1} \forall 1 \leq i \leq m$. The ascending order of elements in a sequence is decided by using the time as the key.

Given a data base $D = \{S_1, S_2, \dots, S_m\}$ that contains m mobile transaction sequences, the support of sequence S is defined as

$$\mathrm{sup}(S) = \frac{|\{S_i | S \subset S_i, 1 \leq i \leq m\}|}{m} \qquad (3)$$

$S = <(t_1, l_1, s_1), (t_2, l_2, s_2), \dots, (t_m, l_m, s_m)>$ is called a frequent mobile transaction sequence if sup(S) is greater than or equal to a specified support threshold δ, and the corresponding TSP is written as $P = TS_i{:}<(l_1, s_1) \xrightarrow[p_1]{} (l_2, s_2,) \dots \xrightarrow[p_{r1}]{} (l_r, s_r)>$, where TS_i semantically represents the time interval between $t_1$and t_r and p_i represents the moving path. With the above definitions, the problem of mining TSPs is defined as follows. Given a database D containing the mobile transactions of users and a specified support threshold δ, the problem is to discover all the TSPs existing in this database.

The spatio-temporal sequence and mobile sequence transaction are mined using Association rule mining. An Association Rule Mining is done on any transaction database. A transaction database TDB is a set of transactions, where each transaction, denoted as a tuple <Tid, X >, contains a set of items (i.e., X) and is associated with a unique transaction identity Tid. A transaction <Tid, X > is said to contain itemset Y if $Y \subset X$. The number of transactions in TDB containing itemset Y is called the support of itemset, denoted as $S(Y)$. Given a minimum support threshold $minsup$and a minimum confidence threshold min_conf, $A \Rightarrow B$ is an association rule if $(A \cup B) \geq min\_sup$ and $\frac{\mathrm{Sup}(A \cup B)}{\mathrm{Sup}(A)} \geq min\_conf$, where A, Bdenote two non-overlapped itemsets, A is called the antecedent, and B is called the consequent.

Most of the traditional association rule mining algorithms divide the mining procedure into two stages. In the first stage, all frequent itemsets are found from the transaction data base. In the second stage, the rules are generated from the frequent itemsets and their confidences are calculated.

3.2 Behavior Analysis

Behavior analysis [7] uses both context and interaction data to analyze the behavior of the user. Context logs collect the history context data and interaction records of mobile users, and thus can be used as data sources for mining behavior patterns.

Given a contextual feature set $F = \{f_1, f_2, \dots,\}$, a context C_i is a group of contextual feature-value pairs, i.e., $C_i = \{(x_1 : v_1), (x_2 : v_2), \dots, (x_l : v_l)\}$, where $x_n \in F$ and v_n is the value for x_n $(1 \leq n \leq l)$. A context with l contextual feature-value pairs is called a l-context.

Contexts may have different granularity with respect to the numbers of contextual feature-value pairs they contain. For example, {(Is a holiday?: No),(Time range: AM8:00-9:00)} is a context with two contextual feature-value pairs, and {(Is a holiday?: No),(Time range: AM8:00-9:00),(Transportation: On vehicle)} is a context with three contextual feature-value pairs so it expresses richer context information than the previous one.

An interaction record is an item in the interaction set $I = \{I_1, I_2, \dots,\}$, where $I_n (1 \leq n \leq Q)$ denotes a kind of user interaction.

Interaction records capture the occurrences of user interactions with mobile devices, such as listening to music, message session or Web browsing. It is also possible to define abstract user interactions, such as entertainment interactions, business interactions, to capture the high level semantic information of user behaviors.

Similar to location prediction and location based service; behavior analysis uses Association Rule Mining for finding interesting patterns from the context logs.

3.3 Intrusion Detection

In intrusion detection, based on the data that is used for the analysis it is classified in to two types as network based monitoring and host based monitoring. Network based monitoring method uses the data that is collected by monitoring the network traffic, data packets, etc. Collecting network based data reduces the overload in the mobile phones and detects external intrusions. As this data does not have information regarding the mobile device it cannot detect threats like malware and Trojan horse. It is also difficult to collect all the network based activity of the mobile device. Host based monitoring method access the data from the mobile phones directly. This helps in accurately identifying the behavior of the device. As this method has direct access to the data in the mobile phones it creates confidentiality issues and also affects the processing ability of the mobile phone. The data mining method that are used for intrusion detection are Clustering and Support Vector Machine (SVM) Classifier.

Clustering is the process of grouping data instances in such a way that similar instances are grouped together, while different data instances belong to different groups. Formally, the clustering structure is represented as a set of subsets $C = C_1, \dots, C_k$ of S, such that: $S = S \cup_{i=1}^{k} C_i$ and $C_i \cap C_j = \emptyset$ for $i \neq j$. Consequently, any instance in S belongs to exactly one and only one subset.

Since clustering is the grouping of similar instances, a measure is needed to determine whether the two instances are similar or dissimilar. There are two type of measures used to estimate the closeness of two instance: distance measures and similarity measures. Distance measure is the most commonly used measure to determine the similarity or dissimilarity between instances. It is useful to denote the distance between two instances x_i and x_j as: $d(x_i,x_j)$. A distance measure is valid only if it is symmetric and obtains its minimum value (usually zero) in case of identical instances. The distance measure is called a metric distance measure if it also satisfies the following properties:

1. Triangle inequality $d(x_i,x_k) \leq d(x_i,x_j) + d(x_j,x_k) \; \forall x_i,x_j,x_k \in S$.

2. $d(x_i,x_j)= 0 \Rightarrow x_i= x_j \forall x_i,x_j \in S$.

An alternative concept to that of the distance is the similarity function $s(x_i, x_j)$ that compares the two vectors x_i and x_j. This function should be symmetrical (namely $s(x_i, x_j) = s(x_j, x_i)$) and have a large value when x_i and x_j are somehow "similar" and constitute the largest value for identical vectors. A similarity function where the target range is [0,1] is called a dichotomous similarity function.

SVM is a supervised machine learning algorithm which uses model for pattern recognition. SVM training algorithm is given a set of training examples, in which all the data belong to any one of two categories, from which models are built that assigns new examples into the appropriate category. An SVM model represents the data from the training set as points in space, mapped in a way that the data from different category are separated by a clear wide gap. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

A SVM is a learning machine that classifies an input vector X using the decision function: f (X) = <X, W>+ b. SVMs are hyper plane classifiers and work by determining which side of the hyper plane X lies. In the above formula, the hyper plane is perpendicular to W and at a distance b / ||W|| from the origin. SVM maximize the margin around the separating hyper plane. The decision function is fully specified by a subset of training samples. This subset of vectors is called the support vectors shown in Figure 3

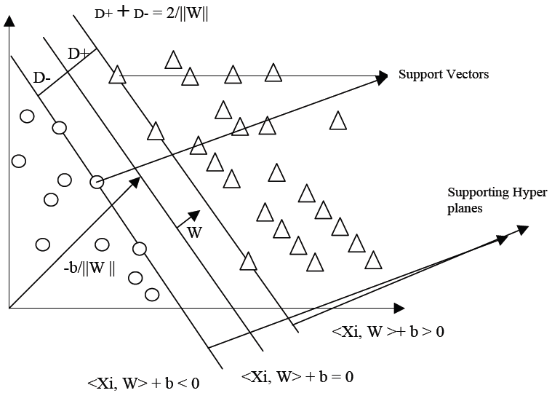

Fig 3: Support Vectors lying on supporting hyper planes

4. CONCLUSION

Mobile data mining is a fast growing area of data mining which gives importance to the mobile phone user. The data collected from the mobile phones give invaluable knowledge representing various aspects of the user thus enabling the vendor to customize mobile usage as per the user needs. In spite of the customization provided to the user the privacy of the user is compromised. Future works can focus on providing user specific functions without compromising the privacy of the user.

5. REFERENCES

[1] NeelamadhabPadhy, Pragnyaban Mishra, RasmitaPanigrahi. 2012. The Survey of Data Mining Applications and Feature Scope. In International Journal of Computer Science, Engineering and Information Technology, Vol.2, No.3, 43-58.

[2] Le-Hung Tran, Michele Catasta, Luke K. McDowell, Karl Aberer. 2012. Next Place Prediction using Mobile Data. In Mobile Data Challenge 2012 (by Nokia), Newcastle.

[3] Jingjing Wang, BhaskarPrabhala. 2012. Periodicity Based Next Place Prediction. In: Mobile Data Challenge 2012 (by Nokia), Newcastle.

[4] Seyed Hasan Mortazavi Zarch, Farhad Jalilzadeh, Madihesadat Yazdanivaghef. 2012. Data Mining For Intrusion Detection in Mobile Systems. In IOSR Journal of Computer Engineerin, Vol. 6, 42-47.

[5] Eric Hsueh-Chan Lu, Vincent S. Tseng, Philip S. Yu. 2011. Mining Cluster-Based Temporal Mobile Sequential Patterns in Location-Based Service Environments. In IEEE Transactions On Knowledge And Data Engineering, Vol. 23, 914-925.

[6] GokulChittaranjan, Jan Blom, DanielGatica-Perez. 2011. Mining large-scale smartphone data for personality studies. In Personal and Ubiquitous Computing.

[7] H. Cao, T. Bao, Q. Yang, E. Chen, and J. Tian. 2010. An effective approach for mining mobile user habits. In CIKM'10, 1677–1680.

[8] Bharat Kumar Addagada. 2010. Intrusion Detection in Mobile Phone Systems Using Data Mining Techniques. M.Sc. Thesis, Iowa State University, Ames, Iowa.

[9] KoenSmets, Brigitte Verdonk, Elsa M. Jordaan. 2009. Discovering Novelty in Spatio/Temporal Data Using One-Class Support Vector Machines. In IJCNN.

[10] Thi Hong Nhan Vu, Jun Wook Lee, and Keun Ho Ryu. 2008. Spatiotemporal Pattern Mining Technique for Location-Based Service System, In ETRI Journal, Vol. 30, 421-431.

[11] GyozoGidofalvi. 2007. Spatio–Temporal Data Mining for Location–Based Services, Ph.D. Thesis, Aalborg University, Denmark.

[12] Fosca Giannotti, Mirco Nanni, Fabio Pinelli , Dino Pedreschi. 2007. Trajectory pattern mining. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining.

[13] GokhanYavas, DimitriosKatsaros, OzgurUlusoy, YannisManolopoulos. 2005. A data mining approach for location prediction in mobile environments. In Data & Knowledge Engineering, Elsevier, 121-146.

[14] Qiankun Zhao, Sourav S. Bhowmick. 2003. Association Rule Mining: A Survey. In Technical Report, CAIS, Nanyang Technological University.

[15] Jae Du Chung, Ok Hyun Paek, JunWook Lee, Keun Ho Ryu. 2002. Temporal Pattern Mining of Moving Objects for Location-Based Service. In LNCS Springer, 331-340.

[16] RakeshAgrawal, RamakrishnanSrikant. 1995. Mining Sequential Patterns. In ICIDE, 3-14.