USERS' NEXT LOCATION RECOMMENDATION USING DATA MINING

[1]Anisha Sharma,[2]Sanjeev Ghosh

[1]PG Student,[2]Associate Professor
[1]Electronics and Telecommunication Department,
[1]Thakur College of Engineering and Technology, Mumbai,India

Abstract :Communication Devices, wireless and web services have allowed mobile users to demand various kind of services on their devices anytime anywhere. Helping users obtain required information effectively is a measure of the quality of service. Predicting the next location of the mobile user as it moves through wireless network plays a key role in maintaining the context-aware services. Previous work done in this field was based on rule-based trajectory mining. For the better recommendation the trajectories are needed to be updated automatically as per the changes in the behavior of the user. This approach consists of a classification-based machine learning algorithm Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel. It gives an accuracy of 6-10% more than traditional methodsof manually updating the system and a good precision stating the relevancy of the system.

***IndexTerms* - Recommendation system, Data Mining, Artificial Intelligence, Location prediction,Machine Learning, Mobile Computing, Mobility Pattern, Classification.**

I. Introduction

The world has experienced a shockwave after the invention of a wireless Personal Communication System(PCS)- Mobile Phones. The name is derived from the term mobility. All thanks to recent advancement in the computer hardware technology which made possible to build small communication devices. PCS supports a huge user population it gives access to such as video, voice, and image at the same time band at any location. The mobility of the users in PCSs gives rise to the problem of mobility management [1]. Mobility management in mobile computing environments covers the methods for storing and updating the location information of mobile users who are served by the system. The focus point of mobility management is mobility prediction.

Human behavior is important to be understood and they are predictable. Demographic features play important role. In the new era of Artificial Intelligence and technology, this can be done by the machine learning and data mining techniques. The feature selection and classification of pattern using Support Vector Machine (SVM) or neural networks can be applied instead of creating trajectory rules as it requires prior information and will recommend on the bases of sequence and changes in the system. New sequences are not recorded and need manual updating. Due to the capability of machine learning classification algorithms the recommendation changes according to the changes in the travelling pattern in user. Python is the high-level object-oriented language is used for the system as it is a scripting language which is easy to understand and is flexible for the integration and the development. It contains the libraries to perform specific set of function. It can run on both Central Processing Unit (CPU) and Graphics ProcessingUnit (GPU) the computational time and speed will differ accordingly.

In the following sections, section II discusses about the related work done in human movement patterns and predicting methods. Section III describes the system design and the steps for implementation of SVM algorithm with RBF kernel is also shown. Section IV highlights the accuracy of the system for daily user, new user and tourist. Section V gives the conclusion followed by future scope.

II. Related Work

A. Human Movement Patterns

Song et al. [2] found that 93% of human movement is foreseeable; The predictability of the individuals' movements is dependent on the entropy of his pattern. It is also shown that whether the individuals' life was within the boundary of 10-km neighborhood or if he travels hundreds of kilometers every day. Lau et al. [3] found the tourist activities as first-time visitor and second time visitor. The study was done for Hong Kong tourism. It is shown in the survey that first time visitors generally follow their itinerary and explore places around their environment bubble in essence closer to their stay. They usually visit sightseeing and more of the shopping than parks or restaurants. Whereas second time visitors have more diverse pattern. Zheng et al.[4] the application is reported to be helpful for tourists and they can discover the landmarks and popular directions easily. It has collected data from the Global Positioning System (GPS). The data contains interesting locations based on the footfall of the individual visitors to find the area of interest.

B. Predicting Methods

The algorithm by G. Yavas et. al. [1] foresees the succeeding inter - cell mobility of the individuals 'mobile. The mobility rules extraction is mined from the past pattern trajectories. The location prophecy is based on such rules. Cannot predict, suggest and store the new trajectory rule. No records of time stamping. D. Katsaros et. al.[5] presented a method called Dynamic Clustering based Prediction (DCP). DCP first collects the mobility patterns from the recorded mobile trajectories. The mobile patterns are discovered then, the prediction of dynamic allocation resource and movements are predicted with the help of these patterns. Weighted edit distance measures in T.lui and P.Behel [6] is used for determining the similarities between two trajectories.T. Anagnostopoulos [7], the author divided into two classifiers – non-perimetric classifier and non-metric classifiers. Non-perimetric classifiers used k-Nearest Neighbour (k-NN) as trajectory classifier. Non-metric classifiers use Decision Tree (DT) as trajectory classifier. It's a cell-based approach. Hidden Markov Model (HMM) cannot be used due to time and space complexity. Non- metric classifier Decision tree gave batter accuracy of prediction and efficiency of predication process.

Ashbrook and Starner [8] various orders of Markov models are applied on the raw data are extracted from the GPS to calculate the probability of the transitions between location. The discussion of the model is qualitative. Prabhala et al. [9] uses Mobile Data Challenge (MDC) data set from Nokia and Week ToDate (WTD) data set from University of California at San Diego (UCSD) for study. The motive was to find the users' next location based on the current location. The algorithm and the model focus on the periodicity of the users' movements. It gives the prediction of next location based on the current location. The program development is done in MATLAB. SVM classification model is used with the help of RBF kernel. The model gave the accuracy of 50%. This amount of accuracy is does not make the model reliable also it does not contain Indian data.

III. System Design

A. Generalized Diagram

The workflow of the system is split into 3 phases: Data Collection Phase, Mining Phase and Learning Phase. As the property of the mobile communication system is diverse, there are different types of data in a log so for the efficient access data is collected and integrated into one data set [10].

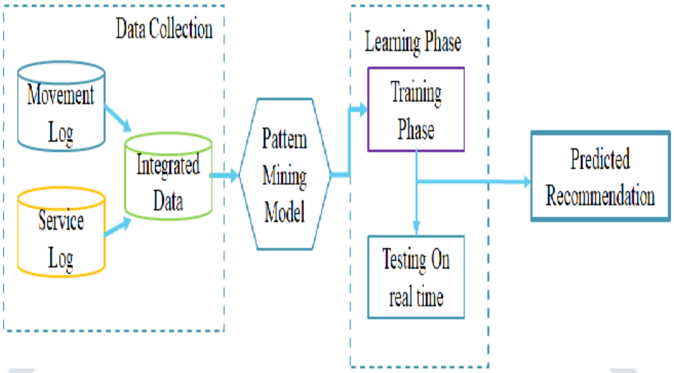


Fig. 1. System Design

The integrated data of movement log and the service log is collected and sent into the model-based data mining technique for the classification. The paths of the pattern are learnt by SVM classifier. For the Learning Phase, the classified data is trained and tested on the real time. To suggest the location the model matches the path from SVM algorithm and predicts the most visited location in the time frame or at the particular sight spot.

B. Algorithm

A SVM is a discriminatory classifier correctly demarcated by a separating hyperplane [11]. In essence, labeled training data, the algorithm yields a finest hyper plane which classifies examples into different classes. The word "supervised" in SVM algorithm because in this work we have feature and class both labels so supervised machine learning for the outcomes. There are two types in SVM classifier—Linear Classifier and Non- Linear Classifier. To find a nearest boundary between the probable outputs- a method called the kernel is used to change the data in required form.

```
Algorithm- Next location Prediction
Input: current place pc,current time interval pt, profile of
          user, most visit ttest , {distpi→pj},pi,pj
          could be any places the user has visited
Output: next place → pp
    1. if there exists a place pn, s.t.
     confidence (pc →pn) >ttest then
    2.        pp = pn;
    3. else
    4      . pp = (sim(dist(pt,),dist(pc →pj)));
    5. return pp;
```

C. Radial Basis Function (RBF) Kernel

Linear hyper plane separation methods are applicable only when the classes are linearly separable. What if one of the classes is on the other side of the hyper plane. At this scenario separation of the classes through division fails. A Standard SVM separates positive and negative example and this is lead to poorly fit model. In most real-life challenges, the data is randomly distributed. Vapnik and Cortes [12] felt the need of a technique which can separate these random data by using a linear classifier to separate a nonlinear problem. The method they introduce is famously known as "Kernel trick".

Two parameters that are considered: C and gamma while training SVM with RBF. The parameter C, is famously known as "soft margin" it allows the SVM to "ignore" some examples for a better overall fit. The C parameter is same in all SVM kernels, there is a trade-off between error penalty and stability. The C indicates SVM on how much misclassification is to avoid on each training data. If the decision function gives the better classification of all the points correctly a smaller margin is considered i.e., value of C will be large. A lower C in decimals will consider a larger margin, though it gives a simpler decision function, irrespective to the fact that large margin will classify the wrong data point. In other words, ``C`` behaves as a regularization parameter in the SVM.

$$k(x,x') = e^{(-\gamma\|x-x'\|^2)} \qquad (1)$$

Equation (1), gamma parameter is the inverse of standard deviation known as the similarity function between the two points. The two points are x is the target element and x^' nearest landmark. If the value of γ is high that means the single training example influence is close to the sample selected as supporting vectors. If the value is low the influence is far from the target. In most of the cases the value for gamma is auto as an indication that no explicit value of gamma was passed.

IV. Results and Discussion

A. Data and Source of Data

The database contains the data about 8 locations that user has travelled to in a span of 2 months. Locations are: office, home, three restaurants, two gardens and one gym. The data generally focuses on the data near to users' home and office in essence the locations he/she visits almost daily depending upon the time. After few days the application starts suggesting location to user at the time. For new user there are total 595 entries and 20 locations near users' area. As the user is new, the application asks to give the interest, age group, gender and marital status. Nine types of interests included: Art gallery, cinema and shopping, gym, library, natural trails, night clubs, parks, restaurants and sports club. The demographic data of the people in the age group of 19-90 years is considered. User has to fill necessary information and on bases of age and interests' places are suggest where most of its age group visits. The user data in both cases are Indians. Tourist's data has 1522 location entries. Sightseeing, restaurants, shopping, hotels, flea market and parks are given as the interests' option. Based on the location he/she wants to visit and interest it will suggest the place.

B. Accuracy

Accuracy is a measure that assures that how reliable the system is. The accuracy is dependent on the verity and volume of the data that is used. Below is the graphical representation of the daily user, new user and the tourist system accuracy. The dotted line represents the hyperplane and the output is circled in the red. The x is the target and x^'is the landmark. One key point is 0 <k(x_,x^') ⩽ 1. 0 and 1 are denoted as the predictions. If -γ(‖ x-x^' ‖)^2~0 that means the point is closer to prediction 0 and if --γ(‖ x-x^' ‖)^2~1 that means the point is closer to the prediction 1.

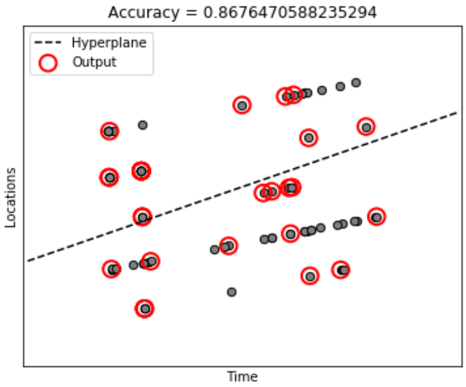


Fig. 2. System accuracy all location

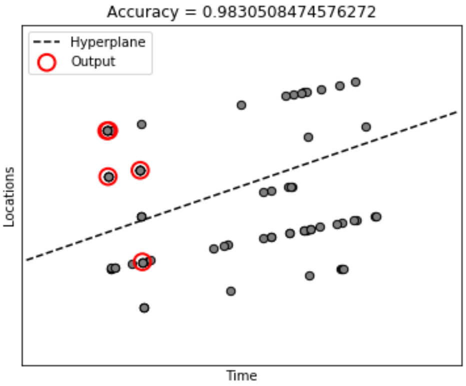


Fig. 3. System accuracy predicted location

Due to the privacy of users' data it was difficult to find the real time data of many users. For the proposed system the single users' activities are monitored for two months. The initial activities are collected from GPS i.e., the latitude and longitude of the location are noted. Also, the time stamp of the user is noted for the temporal based prediction. This information is stored in Standardized Query Language (SQL) database. The integration of the database and Python backend is done by linking it with JavaScript and making server requests in Cross-Platform (X), Apache (A), MySQL (M), PHP (P) and Perl (P)(XAMPP)software. In the fig. 2 and fig. 3 shows the graph with the hyperplane and the output circled in red. The x-axis is denoted as time and y- axis is the location.

Fig 1 shows the possible destinations the user can visit if he/she can visit on that particular day (snap was taken on Monday). Fig 2 represents the accuracy of the place he will visit if user leaves the location A by the particular time (snap was taken at 13:00, Monday). The red circle output to the topmost left will be the least next suggested location where as the others are likely to be next suggested location.

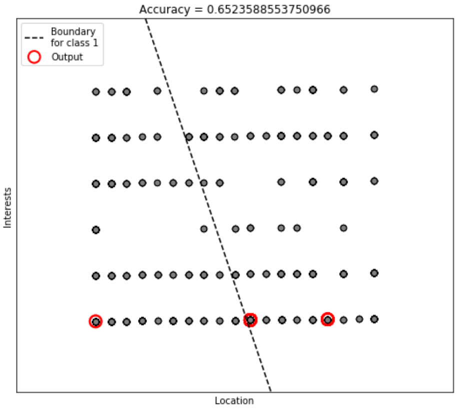


Fig. 4. Accuracy of tourist system

Due to QoS agreement service providers do not share the information of their consumers. The data is built on the bases of reviews and footfall information available online open source and web scraping. As on the bases of the study done in it was seen that generally tourist spend their first few days of trips exploring near their stay. The data was generated by keeping all the parameters in mind. It is easier to number the interests instead of naming them. In the front end the names will be displayed Location parameters of most visited places of sightseeing, restaurant, flea market, shopping, hotels and parks are stored in the database. User can select any one of these at a time. The interests plotted in the y-axis and on the x-axis, locations are plotted as the SVM do not understand the words and the interests are denoted by the numbers.

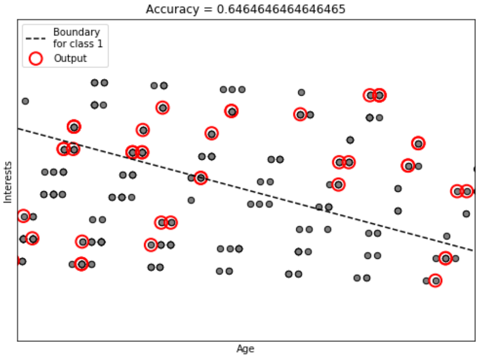


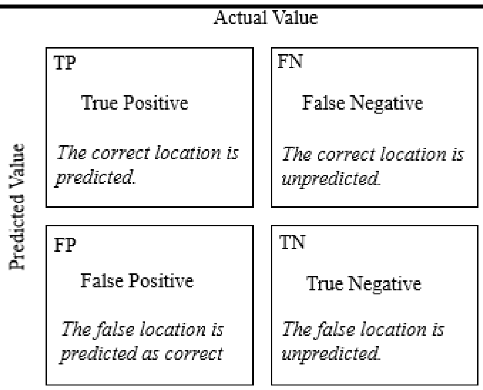
Fig. 5. Accuracy of new user data

New user can also include their activity manually. For a new user in the system age and interest are taken as the parameter of consideration for the suggestion of the next location according to their interest and age group. A survey was conducted with the age group between 17-85 years and with the given set of interests. Their marital status and occupation were also asked to determine if they can be used as a deciding factor. As this also work on linear data the hyperplane parameters are chosen the same way as of tourist. As shown in fig 4 the system accuracy is 65.2 for tourist and fig 5 shows accuracy is 64.6. One of the advantages is that the system gives better accuracy when the data is in more volume or there is an update which was not the case with the rule-based method.

C. Precision and Recall

Precision actually belongs with respect tonext predicted location. It shows how it givemost visited location is actually mostly visited. A precision near to accuracy gives a good performance as the false alarm is minimized. Recall shows the percentage of the truly predicted value over all predicted value. It comes with a disadvantage that it does not take false prediction into the consideration. The confusion matrix in fig. 6 gives a better picture abbreviations are as follows: TP- True Positive, TN- True Negative, FN- False Negative also known as Type I error and FP- False Positive also known as Type II error . Equations (2) and (3) are the formulas to calculate precision and recall. As discussed in the previous subsection that the section of gamma and C is really important. Different combinations of gamma and C gives the different precision and recall for the system. A system with the high precision and low recall gives fewer results but most of the predicted results are correct when compared with training labels. When the system has low precision and high recall it is the opposite, most of the predicted results are incorrect compared to training dataset.

In python the library name scikit-learn also known as ‘sklearn’ has the function of precision and recall. Generally, it is called with the metrics function which is important to be defined earlier. To remove precision and recall one has to define SVM, metrics, from sklearn, train test split model is selected. From metrices matplotlib and precision recall curve is called from matrices. The model is divided into training data of 70% and testing data of 30%. The Support Vector classifier calls SVM classifier from within, the kernel, gamma and C is set and the model is fitted in the classifier.



	Actual Value	
Predicted Value	TP True Positive *The correct location is predicted.*	FN False Negative *The correct location is unpredicted.*
	FP False Positive *The false location is predicted as correct*	TN True Negative *The false location is unpredicted.*

Fig. 6. Confusion Matrix for the proposed system

$$Precision = \frac{TP}{TP + FP} \qquad (2)$$

$$Recall = \frac{TP}{TP + FN} \qquad (3)$$

Table 1: Precision and recall of the proposed system

	Tourist	**New User**
Precision	0.64	0.54
Recall	0.47	0.43

Table 1 indicates the precision and recall of the proposed system for the new user and tourist.In any system it is not possible to increase both the precision and recall. It comes with the trade off to one and other. It can be possible that precision and recall are same. In such a case TP+FN=TP+FP i.e FN=FP. In such case one should check the F1 score or the accuracy.

V. Conclusion and Future Work

The growing use of context-aware devices has directed to an increasing accessibility of trajectory data representing the movement of moving objects. The potential of these data in solving important research problems has raised researchers' interest in analysis methods for them. In this system spatio-temporal data is considered. The accuracy of the single user of proposed system is 83% for the current location and while suggesting the next location on bases of its activities it gives around 98%. A qualitative research done by [8] does not mention the overall accuracy. Also, they have described potential system for multi-user. the proposed system demonstrated the system accuracy of multiple user (new user and tourist) between 64 and 65% respectively. Overall accuracy of system will depend upon the variation of data. For the future work one can use voice-based platforms to know their activity. Integrate the system with the calendar or the events on social media platform to pick up location, time and date for the event. User data is sensitive and prone to leak data security for such a data is a must. There is a scope on increasing the accuracy of the overall system get the data of multiple daily user and calculate the f1 score for the system.

References

[1] G. Yavas, D. Katsaros, OzgurUlusoy, and Y. Manolopoulos, 'A data mining approach for location prediction in mobile environments', Data Knowl. Eng., vol. 54, no. 2, pp. 121–146, 2005.

[2] C. Song, 'Limits of Predictability in Human Mobility Limits of Predictability in Human Mobility', no. February 2010, 2014.

[3] G. Lau, B. Mckercher, G. Lau, and B. Mckercher, 'Tourism and Hospitality Research Understanding tourist movement patterns in a destination : A GIS approach', 2006.

[4] Y. Zheng, L. Zhang, X. Xie, and W. Ma, 'Mining Interesting Locations and Travel Sequences from GPS Trajectories', no. 49, 2009.

[5] D. Katsaros, A. Nanopoulos, and M. Karakaya, 'Clustering Mobile Trajectories for Resource', Springer-Verlag Berlin Heidelb. 2003, no. 102, pp. 319–329, 2003.

[6] T. Liu, P. Bahl, and I. Chlamtac, 'Mobility modeling, location tracking, and trajectory prediction in wireless ATM networks', IEEE J. Sel. Areas Commun., vol. 16, no. 6, pp. 922–935, 1998.

[7] T. Anagnostopoulos, C. Anagnostopoulos, and S. Hadjiefthymiades, 'Efficient location prediction in mobile cellular networks', *Int. J. Wirel. Inf. Networks*, vol. 19, no. 2, pp. 97–111, 2012.

[8] D. Ashbrook and T. Starner, 'Learning Significant Locations and Predicting User Movement with GPS', 2002.

[9] B. Prabhala, J. Wang, B. Deb, T. La Porta, and J. Han, 'Leveraging Periodicity in Human Mobility for Next Place Prediction', vol. 3, pp. 2665–2670, 2014.

[10] V. S. Tseng and K. W. Lin, 'Efficient mining and prediction of user behavior patterns in mobile web systems', Inf. Softw. Technol., vol. 48, no. 6, pp. 357–369, 2006.

[11] F. Liu, D. Janssens, G. Wets, and M. Cools, 'Annotating mobile phone location data with activity purposes using machine learning algorithms', Expert Syst. Appl., vol. 40, no. 8, pp. 3299–3311, 2013.

[12] C. Cortes and V. Vapnik, 'Support-Vector Networks', vol. 297, pp. 273–297, 1995.