ijecs
open access
www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 3 March 2017, Page No. 20801-20805
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i3.73

Cluster Based Data Mining Technique for Identification of User Behavior

J.Kumaran Kumar, E.Karunakaran[2], K.M.Sabarivelan[3]

[1]Pondicherry Engineering College,
Pillaichavady, Puducherry-605014.
kumaran@pec.edu
[2]Pondicherry Engineering College,
Pillaichavady, Puducherry-605014.
ekaruna@pec.edu
[3]Pondicherry Engineering College,
Pillaichavady, Puducherry-605014.
sabarivelankm@gmail.com

Abstract: ***A frameworks for identifying patterns and regularities in the pseudo anonymized Call Data Records (CDR) pertaining a generic subscriber of a mobile operator. We face the challenging task of automatically deriving meaningful information from the available data, by using an unsupervised procedure of cluster analysis and without including in the model any a priori knowledge on the applicative context. Clusters mining results are employed for understanding users' habits and to draw their characterizing profiles. A novel system for clusters and knowledge discovery called LD-ABCD, capable of retrieving clusters and, at the same time, to automatically discover for each returned cluster the most appropriate dissimilarity measure (local metric).The PROCLUS, the well know sub clustering algorithm which is used to identify the sub spaces. The data set under analysis contains records characterized only by few features and consequently to show how to generate additional fields which describe implicit information hidden in data. Also proposed an algorithm over these two techniques for searching common patterns and regularities in order to group together users characterized by a similar profile.***

Keywords: Data mining, Clustering, Patterns, call data records, dissimilarity measures.

1. INTRODUCTION

The development of Information Technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process which knowledge mining from data and knowledge extraction or pattern analysis.

In telecommunication network, Identifying user habits through data mining on call data records [1]. It is a frameworks for identifying patterns and regularities in the pseudo-anonymized Call Data Records (CDR) pertaining a generic subscriber of a mobile operator. The popularity and wide diffusion of cellular phones, a huge quantity of mobile devices are moving everyday with their human companions leaving tracks of theirs movements and their everyday habits. Mobile phones are becoming pervasive in both developed and developing countries and it can be a precious source of data and information, with a significant impact on research in behavioral science.

A Call Data Record (CDR) is a data structure storing relevant information about a given telephonic activity involving an user of a telephonic network. A CDR usually contains spatial and temporal data and it can carry other additional useful information. Population census have been widely used in the past for keeping track of the demography and geographical movements of the population. Nowadays, due to short term and every day mobility, more flexible methods such as various registers and indirect databases are employed: CDRs represent an optimal candidate in this sense. One of their main advantage is that they offer a statistically accurate representation of the distribution of people in an area and they can be used to track large and heterogeneous groups of people.

2. RELATED WORKS

This section contains brief reviews about some existing works in user behavior identification from the call data records.

Aggarwal et al defined a fast algorithms for projected clustering, it is used as a sub clustering algorithm to find out the clusters and the dimensions for the corresponding clusters. It is used to split out those Outliers (points that do not cluster well) from the clusters. It consists of three phases during which the clustering is iteratively improved [2].

Bianchi et al. developed a multi-agent algorithm able to automatically discover relevant regularities in a given dataset, determining at the same time the set of configurations of the adopted parametric dissimilarity measure yielding compact and separated clusters. Each agent operates independently by performing a Markovian random walk on a suitable weighted graph representation of the input dataset. A weighted graph representation is induced by the specific parameter configuration of the dissimilarity measure adopted by the agent, which searches and takes decisions autonomously for one cluster at a time. The results show that the algorithm is able to discover parameter configurations that yield a consistent and interpretable collection of clusters. The algorithm shows comparable performances with other similar state-of-the-art algorithms when facing specific clustering problems. [3].

Ahas et al. proposed a model for the location of meaningful places for mobile telephone users, such as home and work anchor points, using passive mobile positioning data. Passive mobile positioning data is secondary data concerning the location of call activities or handovers in network cells that is automatically stored in the memory of service providers. This data source offers good potential for the monitoring of the geography and mobility of the population. Modeling results were compared with population register data; this revealed that the developed model described the geography of the population relatively well, and can hence be used in geographical and urban studies. This approach also has potential for the development of location based services such as targeting services or geographical infrastructure [4].

Bianchi et al proposed an inexact graph matching algorithm which computes the dissimilarity of a time-varying labeled graph with respect to a static one. This approach is specifically designed for processing very large labeled graphs, which are subject to frequent edit operations that modify the topology and the labeling of restricted zones of the graph. In this scenario, repeating each time an extensive computation of the whole dissimilarity value would require too much time; moreover, since only a specific part of the graph changes, it would result also in a waste of computations. So a fast approach for computing the graph dissimilarity which exploits the dissimilarity value estimated in the previous time interval and the nature of the observed edit operations. The properties of the proposed approach are evaluated with respect to well-known graph matching algorithms, by simulating the dynamics of the graph [5].

Bereta.M used Face recognition based on local descriptors has been recently recognized as the state-of-the-art design framework for problems of facial identification and verification. Given the diversity of the existing approaches, the main objective of this paper is to present a comprehensive, in-depth comparative analysis of the recent face recognition methodologies based on local descriptors. In particular, they highlight the main features in the setting of problems of facial recognition. The presented techniques are particularly suitable for large scale facial authentication systems in which the training stage with the use of the overall face database might be computationally prohibited [6].

3. PROPOSED WORK

The Figure 1 represents framework for the proposed work which is explained as follows: the process of identifying user behavior from the collected input dataset. An Input dataset is sent to data preprocessing which is applied to improve the quality of the input data. The data preprocessing also includes removing user sensitive information if it is there. After performing the data preprocessing the output data is sent to clustering algorithm for identifying the user behavior from the preprocessed call data records. Finally the user behavior and frequent patterns are extracted from the input data. The proposed work will identify the user habit from input dataset, which includes 3 Steps,

3.1 Data Pre-processing

3.2 Clustering

3.3 Generating User Behaviour

3.1 Data Pre-processing

The call data records that have been collected are noisy, thus pre-processing has been applied to improve the accuracy of the input data and to remove the user sensitive information from the call data records. This includes removing of personal information like age, date of birth, gender etc.

3.2 Clustering

In identification of user behavior from call data records the data set need to be clustered accordingly based on certain factors, this can be done by using clustering algorithms. Clustering or cluster analysis is the process of grouping a set of

DOI: 10.18535/ijecs/v6i3.73

objects in such a way that objects in the same group are more similar to each other than to those in the other groups. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields and there are many clustering algorithms are used. The Local Dissimilarities Agent Based Clustering Algorithm (LD-ABCD) and a Projected clustering (PROCLUS) is used for identification of user behaviour from call data records.

A. LD-ABCD

LD-ABCD is a multi-agent algorithm designed to automatically discover relevant regularities in a given dataset, determining at the same time a set of PCs of the adopted parametric dissimilarity measure, yielding compact and separated clusters in the data. Each agent operates independently on a suitable weighted graph, which is used to represent the data. The graph is fully connected, each node corresponds to an element in the dataset and the edges are labeled with a value proportional to the similarity of the two connected elements. The weights on the graph depend on the specific PC m*j* of the dissimilarity measure adopted by the j-th agent. A new PC is iteratively selected by an agent j to construct its own instance Gj of the weighted graph.

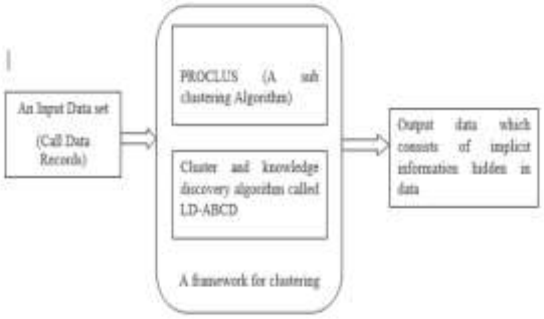
An Input Data set
(Call Data Records)
PROCLUS (A sub clustering Algorithm)
Cluster and knowledge discovery algorithm called LD-ABCD
A framework for clustering
Output data which consists of implicit information hidden in data

Figure 1 Architecture Diagram for proposed work.

B. PROCLUS

PROCLUS is faster than many other subspace clustering algorithms, especially on larger datasets. On the other hand, one of the main drawback of the algorithm is its strong dependence on the parameters k and l which, in many cases, can be hard to be set in advance, since they require an adequate knowledge of the problem and of the dataset at hand. Another drawback is due to the bias toward clusters that are hyper-spherical in shape. Additionally, since the average number of dimensions is given, the number of selected dimension in each cluster will be similar. It is also important to notice that PROCLUS creates a partition of the dataset and, possibly, an additional group of outliers. This means that each instance is assigned to only one cluster (or to the outliers group). This is a critical difference with respect to the procedure implemented in LD-ABCD, which does not form a proper partition, allowing the generation of over-lapping clusters, meaning that an element can be assigned to one cluster, more cluster or no clusters at all.

3.4 Generating User Behaviour

After analyzing the set of CDRs relative to a specific user. Successively, we process the data set without data mining procedure and discussed the results obtained by the two different implementations. The original dataset presented contains CDRs relative to 50,000 users. In this experiments, processed the data relative to the calls of more than 100 different users, which have been randomly selected. For most of them it was possible to identify clear and distinct patterns, while for others we did not obtain meaningful results, mainly because of irregular telephonic activity or for the limited number of the calls issued by the user. In the following, show an example of how an analysis of the CDRs relative to a given user can be performed using the proposed methodology. To visualize the content of the CDRs relative to a given user, there are 4 different charts that show how the CDRs are distributed according to the accounted features. They are:

1. A histogram describing the number of calls done by the user from different prefectures. Each bin is associated with one of the prefectures from where calls were issued and the height of the bars is proportional to the number of calls done in that prefecture.

2. A histogram that describes the distribution of the values contained in the field prev_call. Each bin of the histogram represents the time elapsed from the previous call and its height is proportional to the number of CDRs whose value prev_call falls in that interval.

3. A histogram that represents the distribution of the calls of the user among the 7 days of the week, according to the field weekdays.

4. A histogram that represents the distribution of the calls of the user among the 3 periods of the day, according to the field day period.

Though the data is analyzed on these above mentioned dimensions the group of user that can be identified and the relation between them is identified, but it fails to address that how the relationship between the groups of user is further explored based on the individual relationship with the group of identified user.

Finally, a proposed clustering analysis to analyze the group of identified user for extracting the further more common relations and behaviors of the group of user in the telecommunication based network, this can be done by using one of the clustering algorithm called k-means algorithm for identifying the user behavior for the call data record of a telecommunication network which servers the network to provide services to the each and every user based on their response and interest in calling and also useful for better marketing of the telecommunication network.

4. RESULTS AND DISCUSSION

The results obtained are represented as follows by using the below three graphs.

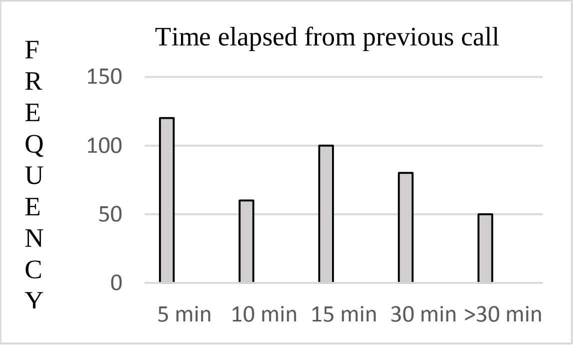

Figure 2 : Time elapsed from previous call

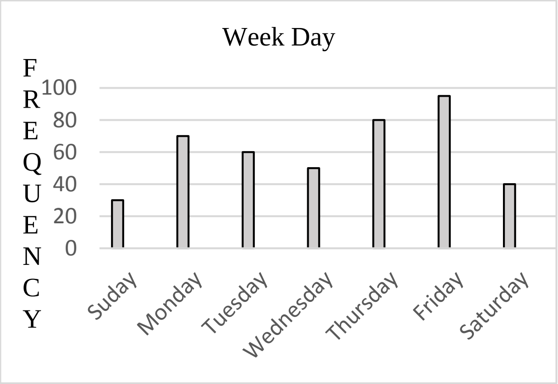

Figure 3 : Week Day

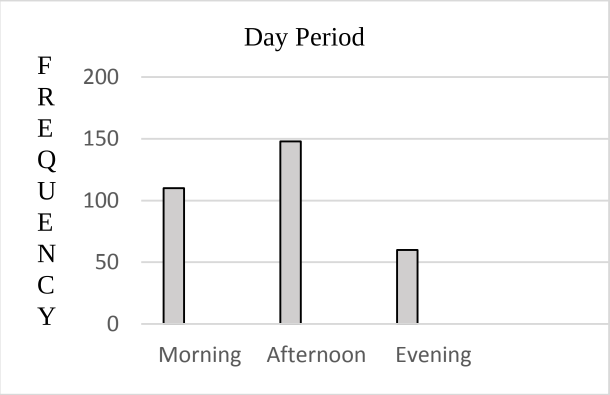

Figure 4 : Day Period

Figure 2 represents the time ealpsed for the previous call, that is time duaration of the previous call palced by the particular user and the graphs shows that the call duration within five minutes is issued more. Figure 3, represents that the call number of calls placed in each and every day of the week that shows that the call is palced more on the last working day of a week i.e Friday. Figure 4, represents the period of time duration in a particular day in which afternoon session has more call than the morining and evening.

Thus evident from the above graphs by using above clustering and sub clustering algorithm the user behaviour, relationships and frequent pattern is identified from the given input set of call data records.

5. CONCLUSION

The proposed work has focused on detecting the user habits from the call data records by using the clustering algorithms, by processing the CDR's of every user in the dataset, searching for common patterns and regularities in order to group together users characterized by similar profile. With the identified cluster of users aim to define specific classes, which can be analyzed and used for describing some general, common behaviors of a customers. The results as above indicates that the proposed work more efficiently identifies the user behavior than that of the existing system.

6. REFERENCES

[1] Filippo Maria Bianchi, Antonello Rizzi, Alireza Sadeghian, Corrado Moiso. " Identifying user habits through data mining on call data records". Elsevier Ltd, 0952-1976, 2016.

[2] Aggarwal,C.C., Wolf,J.L.,Yu,P.S.,Procopiuc,C.,Park.J.S. “Fast algorithms for projected clustering”. SIGMOD Rec.28 (June), 61–72., 1999.

[3] Bianchi,.F., Maior ino,E.,Livi,L., Rizzi,A., Sadeghian,.A. “An agent-based algorithm exploiting multiple local dissimilarities for clusters mining and knowledge discovery”.SoftComput, 1–23, 2015.

[4] Ahas,R., Silm, S.,Jar,O., Saluveer,E., Tiru,M., “Using mobile positioning data to model locations meaningful to users of mobile phones”.J. UrbanTechnol.17 (1), 3–27, 2010.

[5] Bianchi, F.M., Livi, L., Rizzi, A.”Matching of time-varying labeled graphs”. In: The2013International Joint Conference on Neural Networks (IJCNN).pp.1–8, 2013.

[6] Agarwal,R., Gehrke,J., Gunopulos,D., Raghavan,P., “Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications”, ACM,New York,NY,USA,pp.94–105, 1998.

[7] Bereta,M.,Pedrycz,W.,Reformat,”M.Local descriptors and similarity measures for frontal face recognition :a comparative analysis”. J.Vis.Commun. Image Represent.24(8),1213–1231, 2013.

[8] Berlingerio.M., Calabrese,F., DiLorenzo.,G.,Nair, R.,Pinelli,F., Sbodio.,M. “All Aboard: A system for exploring urban mobility and optimizing public transport using cellphone data In: Machine Learning and Knowledge Discovery in Databases”. Springer pp.663-666, 2013.

[9] Bianchi,F.,M., Scadapane,S., Rizi,A., Ucini,A., Sadeghain.,A., “Granular Computing Techniques for Classification and Semantic Characterization of StructuredData”. Cogn.Comput 8(3) 442-461, 2016.

[10] Del vescovo.,G., Rizzi,A. “Automatic classification of graphs by symbolic histograms .In: IEEE International Conference on Granular Computing”, GRC2007.pp.410-410, 2007.

[11] Schliach,J., Otterstatter,T., Friedrich, M. “Generating trajectories from mobile phone data”. In: Proceedings of the 89th Annual Meeting Compendium of Papers, Transportation Research Board of the National Academies, 2010.

[12] Spinosa,E.J.,deLeonF.,deCarvalho,A.P.,Gama,J.a.”Olin dda: A cluster-based approach for detecting novelty and concept drift in datastreams”.In: Proceedings of the ACM Symposiumon Applied Computing. SAC'07. ACM,NewYork,NY,USA,pp.448–452, ,2007.

[13] Japkowicz,N., “Concept-learning in the absence of counter examples”. An auto association based approach for classification,(Ph.D thesis). Rutgers, The State University of New Jerse,1999.

[14] Li, B.,Chang,E., Wu,C-T,.” DPF-A perceptual distance function for image retrieval”. In:Proceedings of the International Conference on Image Processing, vol.2.IEEE,pp.II-597-II-600,2002.

[15] Zhang, S.,Chau,K.-W.,2009.” Dimension reduction using semi-supervised locally linear embedding for plant leaf classification”. In: Emerging Intelligent Computing Technology and Applications. Springer.pp.,948-955,2009.

Author Profile

K.M.Sabarivelan received the B.Tech. Degree in Computer Science and Engineering from Sri Manakula Vianyagar Engineering College in 2015. He is currently pursuing his Master of Technology in Pondicherry Engineering College, (2015-2017) and he is currently doing his Research in the domain of Data Mining.