Wireless Sensor Network, 2012, 4, 113-116
http://dx.doi.org/10.4236/wsn.2012.44015 Published Online April 2012 (http://www.SciRP.org/journal/wsn)

Data Categorization and Noise Analysis in Mobile Communication Using Machine Learning Algorithms

Raghavendra Phani Kumar[1], Malleswara Rao[1], Dsvgk Kaladhar[2]
[1]Department of Electronics and Communication Engineering, GITAM Institute of Technology, GITAM University, Visakhapatnam, India
[2]Department of Bioinformatics, GITAM Institute of Science, GITAM University, Visakhapatnam, India
Email: phanikrch@gitam.edu

Received January 16, 2012; revised February 24, 2012; accepted March 11, 2012

ABSTRACT

Machine learning and pattern recognition contains well-defined algorithms with the help of complex data, provides the accuracy of the traffic levels, heavy traffic hours within a cluster. In this paper the base stations and also the noise levels in the busy hour can be predicted. J48 pruned tree contains 23 nodes with busy traffic hour provided in east Godavari. Signal to noise ratio has been predicted at 55, based on CART results. About 53% instances provided inside the cluster and 47% provided outside the cluster. DBScan clustering provided maximum noise from srikakulam. MOR (Number of originating calls successful) predicted as best associated attribute based on Apriori and Genetic search 12:1 ratio.

Keywords: Traffic; MOR; Data Mining

1. Introduction

The classification (or automated categorization) of texts into predefined categories has spectator with a booming interest in the last 10 years, due to the increased availability of information in digital form in communication technology and the ensuing need to organize them [1]. Technological advances can produce a flood of large data sets that have led to massive data analytic problems and can easily lead to flawed inferences. Statisticians might benefit from learning more about wireless signal controls, and thinking up ways to use data on controls in their analyses [2-4].

Machine learning contains well-defined algorithms, data structures, and theories of learning by automated cauterization or classification of text in to predefined categories [1]. Machine learning became a central research area since mid-1950, due to achieve recognition in artificial intelligence to understand the phenomenon of learning data sets [5].

Pattern recognition and data mining from past few years has fundamental operations in partitioning large set of objects in to homogeneous clusters [6,7]. Scientific data provides a platform to learn the data in search of hidden patterns that exist in large data bases .data mining is the advancement of inductive learning technique to evaluate the usefulness of the cases retrieved from large data sets [8].

In this paper we describe an application of machine learning to an important communication problem: Detection of busy traffic hours in the base stations of an area. We cover the application of machine learning from the formulation of the problem to the delivery of a system for field testing which includes soft handoff traffic and busy traffic hour, soft handoff rate, number of calls, originating calls, paging response, call termination rates. The primary purpose of the paper is to present machine teaching research communities that have general importance in communication technology in machine learning applications.

2. Methodology

The input dataset is in the Waikato Environment for Knowledge Analysis (WEKA) "arff" file format. The Communication data set has 15 attributes and there are 78 instances.

2.1. J48

J48 algorithm is an implementation of the C4.5 decision tree learner, produces decision tree models. The algorithm uses the greedy technique to induce decision trees for classification.

2.2. CART

CART algorithm stands for Classification and Regression Trees algorithm, and is a data exploration and prediction

algorithm. Classification, Regression Trees is a classifier method which in order to construct decision trees.

2.3. SimpleKMeans

In SimpleKMeans clustering, the similarity of two clusters is defined as the similarity of their centroids. The centroid of a cluster which is a point whose parameter values are the mean of the parameter values of all the points in the clusters.

2.4. DBScan

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a data clustering algorithm, as it finds a number of clusters starting from the estimated density distribution of corresponding nodes. It starts with an arbitrary starting point that has not been visited. This retrieves the neighboured clusters, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labelled as noise.

2.5. Apriori

Apriori is a classic algorithm for learning association rules, designed to operate on databases for finding patterns in data.

2.6. Genetic Search

It is a sub set of scoring algorithm, search for multiple solution simultaneously. These solutions blended with each other and are maintained in population based on fitness.

3. Results

The mobile communication depends on the transmitted signal and also the number of users in the cluster. Signal traffic of those mobile users is carried by the base stations. In this paper, the analysis related to the clustering and associated study has been constricted through survey in various base stations, in a clustered area.

J48 pruned tree provided the result with 13 leaves (4 for east Godavari, 2 for Vizag, 4 for Vijayanagaram and 3 for srikakulam). Busy traffic hour has been provided in a leave for east Godavari. The size of the tree contains 23 nodes (**Figure 1**).

CART Decision Tree provides 3 leaf nodes with 5 branches. Signal to noise ratio has been predicted at 55 (**Figure 2**).

SimpleKMeans provided the centroid data for the clustered dataset. About 53% instances provided inside the cluster and 47% provided outside the cluster (**Figure 3**).

DBScan clustering provided maximum noise from srikakulam. Five clusters have been predicted based on the clustering results (**Figure 4**).

```
BTSName = EG
|   SHR2 <= 35.9215
|   |   SHR1 <= 1.1869
|   |   |   BHERL <= 0.1353 : 1 (2.0)
|   |   |   BHERL > 0.1353 : 0 (17.0/2.0)
|   |   SHR1 > 1.1869 : 1 (4.0/1.0)
|   SHR2 > 35.9215 : 1 (3.0)
BTSName = VIZAG
|   SN <= 6.3 : 0 (10.0/2.0)
|   SN > 63 : 1 (5.0)
BTSName = VZM
|   SN <= 42.0 (7.0)
|   SN > 42
|   |   MOR <= 99.1667 : 1 (3.0/1.0)
|   |   MOR > 99.1667
|   |   |   SN2 <= 0.0233 : 0 (4.0/1.0)
|   |   |   SN2 > 0.0222 : 1 (4.0)
BTSName = SKLM
|   SN2 <= 0.0161 : 0 (12.0)
|   SN2 > 0.0161
|   |   SN <= 30:0 (2.0)
|   |   SN > 30 : 1 (5.0)

Number of Leaves :   13

Size of the tree :   23
```

Figure 1. J48 pruned tree.

```
CART Decision Tree

SN < 55.0 : 0 (34.0/11.0)
SN >= 55.0
|   BTSName = (VIZAG) | (VZM) | (EG) : 1 (18.0/8.0)
|   BTSName! = (VIZAG) | (VZM) | (EG) : 0 (7.0/0.0)

Number of Leaf Nodes : 3

Size of the Tree : 5
```

Figure 2. CART decision tree.

The best associated attribute predicted based on Apriori and Genetic search is predicted as MOR (Number of originating calls successful) with 12:1.

4. Discussion

Mobile communication traffic data analysis has been often used as a background application to motivate many data mining problems [9]. The data mining tool tracks for a minimal difference set between things because they believe a list of essential differences is easier to read and understand than detailed descriptions. Summarizing the large data sets to find the data that really matters detailed summaries and generating extensive and lengthy descriptions [10].

A new data mining algorithm which involves incremental mining for user moving patterns in a mobile

computing environment and exploit the mining results to develop data allocation schemes so as to improve the overall performance of a mobile system [11]. Data collected from mobile phones have the potential to provide insight into the relational dynamics of individuals. Distinctive temporal and spatial patterns in their physical proximity and calling patterns allow the prediction of individual-level outcomes such as job satisfaction [12].

Group pattern is used to locate different groups of mobile users associated by means of physical distance and amount of time spent together. Performance of the method indicates a suitable segment size and alpha value needs to be selected to get the best result [13]. Mining frequent sub trees from databases of labelled trees is a new research field that has many practical applications in areas such as computer networks, Web mining, bioinformatics, XML document mining, etc. The application needs more expressive power of labelled trees to capture the complex relations among data entities [14].

Mobile traffic caused by the mobile users in a base station data mining is about finding useful knowledge from the raw data produced by them. Performance evaluation shows that as the number of characteristics increases, the number of rules will increase dramatically and therefore, a careful choosing of only the relevant characteristics to ensure acceptable amount of rules [15].

Attribute	Full Date (78)	0 (37)	1 (41)
SN	48.7308	44.6216	52.439
BTSName	EG	VIZAG	SKLM
CarrierlD	0	0	0
BHERL	1.6716	2.331	1.0765
SH1	0.1048	0.0549	0.1498
SH2	0.5053	0.8845	0.1631
SHR1	5.5352	2.5483	8.2306
SHR2	12.7829	20.1823	6.1054
CN	98.4872	142.6757	58.6098
CON	61.7179	87.5405	38.4146
MO	61.1923	86.8919	38
MOR	97.9378	96.6063	99.1395
PRN	38.1667	58.0811	20.1951
MT	36.6795	55.1081	20.0488
MTR	98.1538	96.9607	99.2306

Time taken to build model (full training data): 0.03 second

=== Model and evaluation on training set ===

Clustered Instances –

0 37 (47%)
1 41 (53%)

Figure 3. SimpleKMeans.

DBScan clustering results

Clustered DataObjects: 78
Number of attributes: 15
Epsilon: 0.9; minPoints: 6
Index: weka.clusteres.forOPTICSAndDBScan.Databases.SequentialDatabase
Distance-type: weka.clusterers.forOPTICSAndDBScan.DataObjects.EuclidianDataObject
Number of generated clusters: 5
Elapsed time: .11

Clustered Instances

0 14 (29%)
1 12 (24%)
2 11 (22%)
3 6 (12%)
4 6 (12%)

Unclustered instances: 29

Figure 4. DBScan.

5. Conclusion

Group pattern of mobile user's results to develop data allocation schemes so as to improve the overall performance of a mobile system without interruption, as the traffic rate is dramatically increasing. Signal to noise ratio has been predicted at 55, based on CART results. The development of intelligent data analysis in mobile communication from the machine learning perspective is necessary in future.

6. Acknowledgements

The authors acknowledged the support from Department of ECE, GITAM University for providing the necessary research facilities.

REFERENCES

[1] Fabrizio Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys* (*CSUR*), Vol. 34, No. 1, 2002, pp. 1-47. doi:10.1145/505282.505283

[2] Leland Wilkinson, "The Future of Statistical Computing," *Technometrics*, Vol. 50, No. 4, 2008, pp. 418-435. doi:10.1198/004017008000000460

[3] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam and E. Cayirci, "Wireless Sensor Networks: A Survey," *Computer Networks*, Vol. 38, No. 4, 2002, pp. 393-422. doi:10.1016/S1389-1286(01)00302-4

[4] D. S. V. G. K. Kaladhar, T. Uma Devi, P. V. Lakshmi, R. Harikrishna Reddy, R. K. SriTeja, V. Ayayangar and P. V. Nageswara Rao, "Analysis of *E. coli* Promoter Regions Using Classification, Association and Clustering Algorithms," *Advances in Intelligent and Soft Computing*, Vol. 132, 2012, pp. 169-177. doi:10.1007/978-3-642-27443-5_20

[5] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, Vol. 1, No. 1, 1986, pp. 81-106. doi:10.1007/BF00116251

[6] M. K. Anderberg, "Cluster Analysis for Applications," Academic Press, Waltham, 1973.

[7] R. S. Michalski and R. E. Strepp, "Automated Construction of Classification: Conceptual Clustering versus Numerical Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 5, No. 4, 1983, pp. 396-410. doi:10.1109/TPAMI.1983.4767409

[8] D. S. V. G. K. Kaladhar and B. Chandana, "Data Mining, Inference and Prediction of Cancer Datasets Using Learning Algorithms," *International Journal of Science and Advanced Technology*, Vol. 1, No. 3, 2011, pp. 68-77.

[9] T. J. Wang, B. S. Yang, J. Gao, D. Q. Yang, S. W. Tang, H. Y. Wu, K. D. Liu and J. Pei, "MobileMiner: A Real World Case Study of Data Mining in Mobile Communication," *Proceedings of the* 35*th SIGMOD International Conference on Management of Data*, Rhode Island, 29 June-2 July 2009, pp. 1083-1086. doi:10.1145/1559845.1559988

[10] T. Menzies and Y. Hu, "Data Mining for Very Busy People," *Computer*, Vol. 36, No. 11, 2003, pp. 22-29. doi:10.1109/MC.2003.1244531

[11] W.-C. Peng and M.-S. Chen, "Developing Data Allocation Schemes by Incremental Mining of User Moving Patterns in a Mobile Computing System," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 1, 2003, pp. 70-85. doi:10.1109/TKDE.2003.1161583

[12] N. Eagle, A. Pentland and D. Lazer, "Inferring Friendship Network Structure by Using Mobile Phone Data," *Proceedings of the National Academy of Sciences*, Vol. 106, No. 36, 2009, pp. 15274-15278. doi:10.1073/pnas.0900282106

[13] J. Goh and D. Taniar, "Mining Frequency Pattern from Mobile Users," *Knowledge-Based Intelligent Information and Engineering Systems*, Vol. 3215, 2004, pp. 795-801. doi:10.1007/978-3-540-30134-9_106

[14] Y. Chi, R. R. Muntz, S. Nijssen and J. N. Kok, "Frequent Subtree Mining—An Overview," *Fundamental Informaticae—Advances in Mining Graphs, Trees and Sequences*, Vol. 66, No. 1-2, 2004, pp. 161-198.

[15] J. Y. Goh and D. Taniar, "Mobile Data Mining by Location Dependencies," *Intelligent Data Engineering and Automated Learning*, Vol. 3177, 2004, pp. 225-231.

