Classification and prediction of social attributes By K-Nearest Neighbor Algorithm with Socially-aware wireless networking-A study

Sujatha krishanmoorthy, Han Bingxu, Zhang Huimin, Chai Yunsheng, Cheng Nuo,
Assistant professor, Students
School of Science and Technology
Wenzhou-Kean University Wenzhou, China
sujatha@wku.edu.cn, zhanghui@kean.edu, chaiyu@kean.edu, chengn@kean.edu

Corresponding Author: Sujatha.ssps@gmail.com, sujatha@wku.edu.cn

Abstract. In this paper, we analyze the data which was collected from SIGCOMM 2009, where 76 users were taken part in and traces of Bluetooth encounters, opportunistic messaging, and social profiles of them were collected. The experiment started by conducting data sorting and data cleaning. In data processing, mainly three main types of characteristic data in the dataset: message, proximity and participant were chosen. After that, Hierarchical cluster analysis was applied on the processed data, which can score the selected segments by similarity measurement, and then form and visually describe the hierarchical structure of these selected clusters. T can either be manually set or a machine language algorithm can be used and here the manually number of categories are set to be 3 after applying the fviz_nbclust function for optimal clustering analysis. As this was a study article five different calculation methods were analyzed to accomplish hierarchical clustering: single linkage, complete linkage, median distance method, UPGMA and Centroid Clustering. The comparison is deployed with a simulation tool and the results are plotted. After the clustering algorithms, the participants are divided into three categories, and for each clustering algorithm the divisions are different. In this regard, k-nearest neighbor classification algorithm is used to calculate the error rate of comparison between the original data and the predicted data. To conclude, Our method results in specific label classification and then analyze the strength of nodes and the tightness between nodes via the KNN algorithm.

1. Introduction

Zhiwen, et al. [1] have defined the socially aware computing. Social perception computing is to perceive and identify the behavior of social individuals in real time, analyze and excavate the characteristics and rules of group social interaction, assist individual social behaviors, and support the interaction .communication and collaboration of the community through the large-scale multi-
type sensor devices increasingly deployed in human living space. The core of social perception computing lies in the word "perception", which has two meanings: first perceiving the real world, then perceiving and responding to it. Social perception computing emphasizes the use of advanced computer science and technology to perceive individual behaviors and group interactions in the real world, understand human social activity patterns, and provide intelligent assistance and support for individual and group interactions. [1]. In the Internet era, especially the emergence of Web 2.0, researchers have been provided with massive data sources for analyzing online behaviors and social networks. Socially aware computing has been booming, and social software and applications have been emerging, such as netizens' behavior pattern mining and social software and applications have been emerging, such as netizens' behavior pattern mining and network public opinion analysis. Therefore, Socially aware computing can be used for human behavior and interaction analysis, individual behavior recognition, and analysis of behavior characteristics and rules [2-3]. Based on the data of individual behavior, mobility and social interaction, the features of group social interaction are analyzed by means of social network analysis, machine learning and data mining. Therefore, in this paper has taken advantage of the abundant socially interaction data resources from the internet and classify and predict network users by utilizing the K-Nearest Neighbor Algorithm.

2. Collation and Classification of Data

2.1. A Data cleaning and Sorting

These data were collected in two field experiments using mobiclique smart phones in two large-scale network conferences. They show who each node (personal ID) is associated with, interest classification of each node, regional characteristics, participant characteristics and sent information characteristics.[4] After the data is collected, it needs to be sorted out first. Put the original disorderly data into excel table, which is conducive to observation and calculation. We use R language to clean the data. Data cleansing plays an important role in data management and data analysis. At the same time, data cleansing, cleansing or appending can effectively help the subsequent processing of data, because it can correct or delete inaccurate and damaged data.[5] First, the data is transformed into CSV file type, and then imported into R language. Through STR function, the structure of each group of data is obtained, including the data type and data length of each column vector. And get the network graph for the connection between nodes. A total of 76 nodes, for the ring layout and hierarchical layout drawing corresponding to the network diagram. These network diagrams can let us clearly see the connections of these nodes, and clearly show that there are complex but close connections between each node. This is good for data classification. The network diagram of the ring layout clearly shows that there are different connections between the 76 nodes. The connections between some nodes are closer and more complex, but the connections between some nodes are relatively sparse. This strongly suggests that there are different categories between these relationships, so we can classify these data. In data processing, we choose three main types of characteristic data: message, proximity and participant. We use statistical nodes to audit the quality of data. From the audit results, we find
that there are missing data, and we process the missing data. We choose to delete these missing nodes.

2.2 Hierarchical cluster analysis

After the preliminary data cleaning and sorting, three kinds of main data information can be selected as the eigenvalues, and then the data classification can be completed. Cluster analysis is to divide a group of data into several categories according to similarity and difference. Its purpose is to make the similarity between data belonging to the same category as large as possible, and the similarity between data in different categories as small as possible. We choose hierarchical clustering to analyze, and try to classify them according to the messages, proximity and participants. Hierarchical clustering analysis can score the selected segments by similarity measurement, and then form and visually describe the hierarchical structure of these selected clusters.

Figure 1: Ring layout
Figure 2: Hierarchical layout

Figure 3: hierarchical clustering line graph
This approach allows us to understand the relationship between fragments more easily. However, hierarchical clustering cannot automatically determine the number of categories, it is necessary to define the number of categories of the final clustering in a user-defined way. We can use fviz_nbclust function for optimal clustering analysis.

In the resulting line graph, this clearly shows that the data classification takes a big turn in 2 and 4. But when we grade these data, two categories may be too few and four categories may be too many. Therefore, we finally decided to divide it into three categories. In the preliminary clustering graph, the data are clearly divided into three categories, and only a few outliers exist. We can clearly feel that these data distribution have hierarchical clustering, and further prove that we can divide into three categories. These data will be finally hierarchical clustering through five different calculation methods: single linkage, complete linkage, median distance method, UPGMA and Centroid Clustering. According to these five methods, different types of classification will be further processed by using the model.
The first is the single linkage. Single linkage can process data of various shapes and densities, but it is sensitive to noise. A single linkage considers local density to ensure that the distance involves the core of each group.[7] The single linkage clustering method is to find out the non-diagonal elements of the original $m \times m$ distance matrix, to merge the classification objects GP and GQ into a new class GR, and then according to the calculation formula, calculate the distance between the original classes and the new classes, so as to get a new $(m-1)$ Order distance matrix; then select the smallest $D_IJ$ from the new distance matrix, and merge GI and GJ into the new class; then calculate the distance between the new classes and the new classes, so on until the classification pairs are classified into one class. Through this calculation method, figure 5 can be obtained. Figure 5 clearly shows that tags 1 and 22 are divided into one category, tag 38 is a separate category, and all remaining data tags are a separate category. Then through the artificial custom division, set a, B, C three different levels of categories for the next step model test.

The second is complete linkage. The difference between complete linkage and single linkage is that the formula used to calculate the distance between the original class and the new class is different. The longest distance clustering method uses the longest distance to measure the distance between samples. Some algorithms may have high time complexity. But complete linkage algorithms can provide the best possible result in the manageable time. Compared with the solution of single linkage hierarchical clustering which can find individual clusters with high precision, complete linkage can find the optimal solution according to the linkage between data.[8] In this way, we can get Figure 6. Figure 6 clearly shows that all data can be divided into two categories and one subclass. In addition, we set three different levels of categories A, B and C for the next step of model test through custom division.

The third method is median distance method. Intermediate distance method is a kind of system clustering method. It is a compromise between the minimum distance method and the maximum distance method. For the distance between classes, the minimum distance
between two classes of samples and the maximum distance between two classes of samples shall not be taken, but a certain intermediate distance shall be taken. This method uses the center theorem to find the center of a class, and then finds the minimum distance between centers in different classes. According to the classification diagram obtained in Fig. 7, it can be clearly seen that labels 10 and 46 are in the same category and labels 1 and 22 are in the same category. This is similar to the previous single linkage.

The fourth is UPGMA (unweighted pair group method with arithmetic mean), and the classification result is shown in Figure 8. Average link analysis is usually called UPGMA (unweighted pair group method with arithmetic mean). Because big data sets are becoming more and more popular in the field, and UPGMA is easy to understand conceptually and very fast in practice, it can be said that it is the most popular hierarchical clustering algorithm, and it is still widely used and very popular.[9] The fifth is Centroid Clustering, and the classification result is shown in Figure 9. This method is simple and efficient, which is suitable for both the tasks to be solved and the internal clustering module.[10] There are one or two labels in the two categories as a separate category, which may be caused by the influence of outliers. In Figure 8, tags 1 and 22 are a separate category, and the remaining data are classified into two categories. In Figure 9, label 30 is a separate category, and label 1 and 22 are a separate category. This is similar to the previous classification results.
After getting these detailed hierarchical clustering, hence after customizing the three categories of selection labels according to the attributes of the data analysis and prediction can be performed. Compared with these clustering analysis, these classification charts clearly show that for tags 1 and 22, they have high similarity and are classified into the same category. This means that in the follow-up research, we can better mine the relevance between the two tags. It may even serve as the basis for classification so that when new data is obtained, it can be classified according to these two labels. At the same time, the result graph of clustering based on the longest distance method shows that there is no single or very small amount of data that is clustered into a single group, which may be related to the characteristics of classification, so it needs to be further tested and observed in the analysis and prediction.

3. USING K-NEAREST NEIGHBOR ALGORITHM FOR ANALYSIS AND PREDICTION

After five clustering algorithms, the data set grouping will be obtained, which is composed of many similar objects. Through clustering algorithm to find the internal structure of data, all data instances are organized into some similar groups. The data instances in the same cluster are the same as each other, and the instances in different clusters are different from each other. When using five groups of algorithms, the original data will be divided into three groups. In this regard, k-nearest neighbor classification algorithm is used to calculate the
error rate of comparison between the original data and the predicted data. [11] Combined with the advantages of KNN, the visualization results are constructed and the error rate prediction results are analyzed according to the data. For each node variable, we use the data of N nodes in the dataset. For the given data, the value of n is set to 76. The specific attributes of 76 nodes are copied out to five copies, which are labeled with five groups of feature classification labels analyzed by clustering algorithm before. According to each copy, the text file is generated, parsed by python, and a two-dimensional diffusion graph is drawn by using Matplotlib. Part of the data of the text file is used as the test sample to calculate the error rate. If the error rate is within the acceptable range, k-nearest neighbor algorithm can be run to classify. From the beginning of k = 1, gradually increase, use the test data to analyze the accuracy, so as to choose the best K. This can provide a reference for communication operators to implement the D2D payload balance and scheduling. [12] At the same time, it has some reference significance for ICN. The data traverses the network according to the match between its name and the user's interest in this kind of content, which is independent of the network location and realizes the efficient and rich content transfer. [13]

3.1 Limitations

In data processing, we choose three main types of feature data: messages, proximity and participants. In the analysis of all the overall data, there are some limitations. After using clustering algorithm for analysis, the label results can be verified by k-NN algorithm, and the label category of node attributes can be obtained accurately, so as to infer the strength and weakness of information nodes. In this respect, the strength of information nodes and the compactness of adjacent nodes need more research to prove. When KNN model is used to analyze data, it has the limitation of parameter selection and heavy learning time. [14] Despite these limitations, we are still trying to analyze and predict data results in social networks and communications, which will be discussed below.

3.2 Analysis Process and Results

In the initial analysis, we get and separate the original data set. The value of each node is extracted from the participants data table, and the device_average data is calculated from the proximity data table. Moreover, the corresponding created messages data of each node is extracted from the messages data table. These data are collected and imported into a data table. After the analysis of five clustering algorithms, five different copies of data table are obtained. In these copies, each node has a label attribute generated by clustering algorithm. These data are parsed by python, and two-dimensional diffusion graph is drawn by Matplotlib.
We can get figure 10,11,12 according to the data copy of classification analysis using the shortest distance method. Through the visualization of data, we can clearly feel the decentralized relationship and diffusion form between the two attribute tags. The data analyzed by other four clustering algorithms are also analyzed in the same way. Through comparison, from the perspective of visual results, their distribution pattern has some similarity. In the equation, the attribute with the largest number difference has the greatest impact on the calculation results, that is to say, messages will have a far greater impact on the calculation results than the other two features. The reason for this is that the number of messages is larger than the other two eigenvalues. If these three-feature data are regarded as equally important, we need to balance their feature weights to prevent any of them from seriously affecting the analysis and prediction results. When dealing with the eigenvalues of different value ranges, we use the method of numerical normalization to process the value ranges from 0 to 1. We use the following formula to normalize the three characteristic attributes:

$$\text{NewValue} = \frac{\text{OldValue} - \text{Min}}{\text{Max} - \text{Min}}$$

The normalized data is obtained by subtracting the minimum value from the original value and dividing it by the difference between the maximum value and the minimum value. Moreover, the range of data and the minimum value of data are calculated, and the data is preprocessed. Next, 60% of the selected node attribute data is selected as the training sample to train the classifier, and the rest 40% data is used to test the classifier to detect the accuracy of the classifier. This 40% can be selected randomly from the data to prevent affecting its randomness. We chose the option of including three labels in the measurement to prevent the situation that there are only one or two labels in the test sample. The error rate is 13.043478% by setting the K value to 1 and cross checking the experiment. At the same time, we analyze and test the other four copies, and get their error rate as follows: 34.782609% (K=3), 17.391304% (K=3), 4.347826%. The change of error rate is detected by changing the value of function, inner variable and classifier K. It can be seen from the results of the validation classifier that the error rate of KNN is 4.347826%. Compared with other classification results, it is a result with the lowest cross validation error rate. Through the analysis, we can get that this model has a low error rate for the data classified by barycenter method, which proves the accuracy of the algorithm.
Therefore, the new input data can be predicted, and the expected results can be obtained. According to the prediction results, we can get the specific label classification, and then analyze the strength of nodes and the tightness between nodes.

4. Deployment and application

The purpose of this work is to analyze the high-intensity information nodes and their close relationship by establishing a model. From the expected results with the analysis high-intensity information nodes can be determined. These nodes can be applied in many fields.

4.1 Positioning with nodes.

Positioning with information nodes is very important for wireless networks. Based on the received signal strength, the position estimation of the source node is generally carried out by converting the received signal strength measurements of several known positions into the distance from the source node. [15] The process is as follows: the transmitting node's transmitting signal strength is known, the receiving node calculates the transmission loss of the signal according to the received signal strength, transforms the transmission loss into distance by using theoretical and empirical model, and then calculates the node's position by using the existing algorithm. Through the model, the high-strength nodes are determined, and the distance to the source node is calculated by detecting its strength measurement value, which can help to locate. For some network operators, calculating the location of high-strength nodes and ranging according to their strength can help them determine the location of the signal tower and maximize the communication resources as much as possible. In addition, node location is widely used in military, transportation, industry, and environmental fields. For example, military reconnaissance, radar, intelligent transportation, environmental monitoring, emergency rescue, etc.

4.2 Provides a reference for D2D to achieve payload balance and scheduling.

D2D (device to device) communication is a new technology, which allows terminals to directly communicate by reusing cell resources under the control of the system. It does not rely on the central node, but on the edge nodes of the network, to realize self-organization and peer-to-peer collaboration of resource discovery and sharing. D2D has the potential to improve system performance, enhance user experience and expand the application prospect of cellular communication. D2D adopts the authorized frequency band of telecom operators, with controllable interference environment and high data transmission reliability. In the edge computing network of D2D, the task will choose to compute at the sensor node or on the edge server according to its own attributes and the load balance state of the network. [16] When the edge server is overloaded, it needs to schedule tasks to other relatively idle servers to achieve load balancing between the edge servers. This model infers the load of the task on the node by judging the strength of the node, so as to judge whether the node is overloaded. This provides a reference for D2D to achieve load balancing and scheduling.

4.3 Provide a new link prediction algorithm.

Link prediction is an effective method of data mining and prediction analysis in social networks. Link prediction is mainly divided into two types: one is to predict the future links according to the current network status. Second, it can identify existing but unknown links.
[17] Link prediction plays an important role in some fields, such as information retrieval, system recommendation, etc. In the future, through the model to determine the node strength, take the public neighbor node as the main research object, to study the link strength between each public neighbor node and two neighbor nodes, so as to predict the nodes. The link strength of the common neighbor node and the prediction node will affect the link prediction. Compared with other algorithms for the similarity between nodes, using the link strength of nodes to predict the link has better test results while maintaining the same time complexity.

5. Conclusion

In conclusion, this paper analyzes and processes the sigcom data. Then, five different calculation methods are used to analyze the processed data: single link, full link, median distance, UPGMA and centroid clustering. After scoring each data segment, the hierarchical structure of clustering is described intuitively by image. After hierarchical clustering analysis, k-nearest neighbor algorithm is used to classify and predict network users. After getting the prediction data, compare with the original data, and get the error rate. The result shows that the model has low error rate for the data classified by barycenter method, which proves the correctness of the algorithm. Through this model, we can determine some high-intensity information nodes. In the subsequent work development, the strength of these nodes and their close relationship can be determined. These identified high-intensity information nodes also have important applications in some fields. For example, in the fields of military, industry, communication, and transportation, nodes are used to locate; for D2D, it provides a reference for load balancing and scheduling; for link prediction, it provides a new algorithm.[18-22] There are many different application were the clustering can be applied.

References

[1] Yu Zhiwen, Yu Zhiyong, Zhou Xingshe. Social perception computing: concepts, problems and research progress [J]. Journal of computer science, 2012, 35 (01) : 16-26.
[2] Chen L, Nugent CD, Cook D, Yu Z. Knowledge-driven activity recognition in intelligent environments. Pervasive and Mobile Computing 2011, 7 (3): 285 - 286.
[3] Madan A, Caneel R, Pentland A. GroupMedia: Distributed multimodal interfaces //Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI. State College, Pennsylvania, USA, 2004: 309-316)
[4] A.-K. Pietiläinen, E. Oliver, J. LeBrun, G. Varghese, and C. Diot, "MobiClique: middleware for mobile social networking," in Proceedings of the 2nd ACM workshop on Online social networks, 2009, pp. 49-54.
[5] Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016, June). Data cleaning: Overview and emerging challenges. In Proceedings of the 2016 International Conference on Management
of Data (pp. 2201-2206).

[6] Heckerman, D. E., Bradley, P. S., Chickering, D. M., & Meek, C. A. (2008). U.S. Patent No. 7,333,998. Washington, DC: U.S. Patent and Trademark Office.

[7] Ros, F., & Guillaume, S. (2019). A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise. Expert Systems with Applications, 128, 96-108.

[8] Mamun, A. A., Aseltine, R., & Rajasekaran, S. (2016). Efficient record linkage algorithms using complete linkage clustering. PloS one, 11(4).

[9] Moulton, V., Spillner, A., & Wu, T. (2018). UPGMA and the normalized equidistant minimum evolution problem. Theoretical Computer Science, 721, 1-15.

[10] Huang, G., Larochelle, H., & Lacoste-Julien, S. (2019). Centroid networks for few-shot clustering and unsupervised few-shot classification. arXiv preprint arXiv:1902.08605, 3(7).

[11] J. Shi and L. Yang, "A Climate Classification of China through k-Nearest-Neighbor and Sparse Subspace Representation," Journal of Climate, vol. 33, no. 1, pp. 243-262, 2020.

[12] Y. Cao, C. Long, T. Jiang, and S. Mao, "Share communication and computation resources on mobile devices: A social awareness perspective," IEEE Wireless Communications, vol. 23, no. 4, pp. 52-59, 2016.

[13] W. Moreira and P. Mendes, "Social-aware forwarding in opportunistic wireless networks: Content awareness or obliviousness?," in Proceeding of IEEE international symposium on a world of wireless, mobile and multimedia networks 2014, 2014: IEEE, pp. 1-6.

[14] G. Alimjan, T. Sun, Y. Liang, H. Jumahun, and Y. Guan, "A new technique for remote sensing image classification based on combinatorial algorithm of SVM and KNN," International Journal of Pattern Recognition and Artificial Intelligence, vol. 32, no. 07, p. 1859012, 2018.

[15] H. Wen, Z. Fang, M. Wang. (2017). “A source node location algorithm based on received signal strength.” JOURNAL OF GEODESY AND GEODYNAMICS, vol. 37 Issue (4): 380-384

[16] R Wang, x.nie, d. Wu, h. Li. (2020). "Social attribute aware task scheduling strategy for edge computing. "Journal of electronics & information technology. vol. 42no. 1. P271-278

[17] Y. Guisheng, y. Wansi and d. Yuxin, "A new link prediction algorithm:

[18] Node link strength algorithm," 2014 ieee symposium on computer applications and communications, weilai, 2014, pp. 5-9, doi: 10.1109/scac.2014.8.

[19] Punithavathani, D.S., Sujatha, K. & Jain, J.M. Surveillance of anomaly and misuse in critical networks to counter insider threats using computational intelligence. Cluster Comput 18, 435–451 (2015). https://doi.org/10.1007/s10586-014-0403.

[20] Lakshmanaprabhu, S. K., Mohanty, S. N., Krishnamoorthy, S., Uthayakumar, J., & Shankar, K. (2019). Online clinical decision support system using optimal deep neural networks. Applied Soft Computing, 81, 105487.

[21] Anbarasance, M., Muthu, B., Sivaparthipan, C. B., Sundarasekar, R., Kadry, S., Krishnamoorthy, S., & Dasel, A. A. (2020). Detection of flood disaster system based on IoT, big data and
convolutional deep neural network. Computer Communications, 150, 150-157.

[22] Appathurai, Ahilan, et al. "A study on ECG signal characterization and practical implementation of some ECG characterization techniques." Measurement 147 (2019): 106384.