An OSM Contributors Classification Method Based on WPCA and GMM

Yijiang Zhao, Xingcai Wei, Yizhi Liu and Zhuhua Liao

Key Laboratory of Knowledge Processing and Networked Manufacturing in Hunan Province, Hunan University of Science and Technology, Xiangtan, 411201, China
Email: wxc972547526@163.com

Abstract. Contributors have a significant impact on data quality of OpenStreetMap (OSM) because most of them are the non-professional, so clustering analysis of contributors based on different experiences has practical significance. Firstly, this paper obtained 31 behavioural characteristics of contributors from OSM historical data. Then, a weighted principal component analysis (WPCA) method was used to reduce the dimensions of the contributors’ behaviour in the selected region. By using an unsupervised prototype-based Gaussian mixture model (GMM) clustering algorithm, contributors with similar contribution attributes in the London area were clustered into four groups. Finally, the characteristics of four different types of contributors are analysed, and two types of experienced and professional contributors are found, who contribute a large amount of high-quality data.

Keywords. OpenStreetMap; contributors classification; WPCA; GMM.

1. Introduction

Volunteer Geographic Information (VGI) is a term put forward by Goodchild in 2007. This term appeared in the era of “Web2.0”, in which everyone can spread VGI through Internet technology and innovation. VGI encourages citizens to voluntarily acquire and contribute geographic data by using various tools and methods (handheld GPS devices, aerial photographs, and even local knowledge). At present, VGI data is widely used in road detection [1], urban modeling [2].

OSM is an open-source map project, which provides everyone with a free and open access global geographic feature database [3]. Since its establishment in London in 2004, OSM has attracted more and more people’s attention after years of continuous development. In some respects, it can even replace official data sets and become a potential competitor of public and commercial geographic data providers [4]. Nowadays, OSM has become the most representative project of VGI, and it has increased considerably in terms of data volume and the number of contributors. Heipke pointed out that since the data contributed by contributors are mostly based on their living scenes, they combine with their local knowledge, which further improves the quality of OSM data [5]. As a representative project of VGI, OSM data is based on the collection and contribution of contributors. At the same time, the OSM data contributors are anonymity, some contributors may hardly geospatial information education background and work experience, their contribution to the data may not be able to accurately describe the real world, this leads to data quality is not consistent, so the trustworthiness and reliability of the data quality has become a major problem [6].

At present, there are two methods to evaluate the quality of OSM data, one is the external quality evaluation method, the other is the internal quality evaluation method. Among them, the external quality assessment method is mainly defined by several quality characteristics by comparing with the external
official authoritative datasets. However, access to authoritative data is limited in availability, limited in license, and high in procurement cost. Therefore, more and more people pay attention to the evaluation of internal OSM data quality recently. In 2014, Gröchenig et al. proposed a new method to evaluate the integrity of OSM database by analyzing the changes of volunteer community activities over time, which opened up the research field of internal indicators to evaluate the quality of OSM data [7]. Since the quality of contribution data directly or indirectly affects the quality of the whole OSM data, the contributor plays an important role in the data quality of OSM. Therefore, by classifying OSM contributors, this paper can directly evaluate the quality of OSM data according to the classified categories of contributors.

Haklay proposed to classify the contributors by studying the motivation and participation of volunteers’ contribution data. Budhathoki and Haythornthwaite divide OSM contributors into serious and random mapmakers, mainly considering the number of their contribution objects, the duration, and frequency of their contribution [8]. Jacobs and Mitchell (2020) proposed a method based on unsupervised machine learning to classify contributors and further evaluate data quality [9]. However, after dimensionality reduction using PCA, it is impossible to clearly distinguish each variable, which leads to the inability to distinguish the characteristics of contributors’ contribution habits. Therefore, this paper proposes to use WPCA, by attaching weights to several variables, the main contribution habits of users can be identified more clearly, and the contributors can be better classified.

2. Data Description
The research area of this paper is London, England, and OSM data of London is obtained through Geofabrik. Figure 1 shows the regional scope of London, with its center at 51°30′ N and 0.1°5′E, covering an area of 1577 square kilometers.

![Figure 1. Study area (London).](image)

In this paper, through the framework developed by Oslandia (a company focusing on open-source GIS architecture and software solutions), the contributor metadata information is obtained, including 14902 contributors and 31 variables describing behaviors.

3. Methodology
In this paper, the contribution history characteristics of the contributors in the study area are selected firstly, and then the characteristics of the contributors' metadata elements extracted by Oslandia are weighted and then reduced in dimension by using WPCA, in which the selection of the number of components and the determination of the weights are the key steps. Then, an unsupervised prototype-based GMM is used to cluster contributors with similar contribution attributes into different groups. Finally, the characteristics of different types of contributors are obtained through further analysis.

3.1. Weighted Principal Component Analysis Method
PCA is the principal component analysis method, which is a statistical method. It can be used to analyze
several possible related observations, expose the trends or patterns among variables and reduce redundancy. Let X be a \( d \times n \) dimensional matrix, and then calculate the sample mean in the original data for the data in a d-dimensional, as follows:

\[
\bar{X} = \frac{1}{n} \sum_{k=1}^{n} X_k
\]  

The covariance matrix was calculated as follows:

\[
P = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \bar{X})(X_k - \bar{X})^T
\]

Then, calculated all eigenvalues and eigenvectors in the covariance matrix, of which the original sample is X. The eigenvectors corresponding to the first d ‘maximum eigenvalues are selected to form matrix Y, and finally, the first d ‘maximum principal components constitute matrix B. The calculation is as follows:

\[
B = YX
\]

It analyzes the variance of each potential explanatory variable relative to a response variable and classifies the explanatory variable with detailed covariance as a principal component. The influence of the change of each specific variable on the correlation between each principal component and the response variable is called the loadings of the component. However, the projection data of traditional PCA does not distinguish the class labels of the data, and it is often ambiguous to explain the meaning of principal components after dimension reduction, which makes it difficult to distinguish sample attributes [10]. To address the problems of traditional PCA, this paper uses a weighted principal component analysis (WPCA) method proposed by L. Delchambre, which is optimal for identifying important features in a dataset [11]. Based on preprocessing the original data, an appropriate weight calculation method is selected to weight the creation behaviors of node, way, and relation, and then PCA feature extraction and dimension reduction are performed on the data samples.

3.1.1. Component Selection. There are two methods to select components, the first is to select components to cover 95% variance, and the second is to consider components with eigenvalues greater than 1. However, as shown in figure 2, there are no components with eigenvalues greater than 1, so the second method fails. This paper will use the first method to select 6 components.

![Figure 2](image)

*Figure 2.* Most of the variance can be explained by the first six principal components.

3.1.2. Weight Selection. In this paper, the preprocessed OSM data samples are weighted before feature extraction and classification, which is more beneficial to make the OSM data samples in this paper tend to be classified. In this paper, a weight calculation method proposed by Huang et al. [12], is used, which is simple and effective, and the weight calculation formula is as follows:
where: vector \( c = [1, 2, \ldots, p] \) represents a class, \( S_c = \sum_{c \in C} X_c \) represents the sum of all elements in a class of data, matrix \( X_c \) (\( c \in C \)) represents a subset of samples belonging to class \( c \), \( S_C = \sum_C X_C \) represents the sum of all data elements, matrix \( X_C \) represents all data samples of class, and \( W_C \) is the weight of a class of data. Finally, all the elements of each category are uniformly multiplied by the weights corresponding to the category.

3.2. Gaussian Mixture Model Clustering

GMM is an unsupervised prototype-based clustering algorithm, in this paper, GMM is used to cluster contributors with similar contribution attributes into groups. This clustering algorithm usually assumes that the clustering structure of this paper can be described by a set of prototypes, which is a very common algorithm in the real clustering task.

GMM is to use multiple linear combinations with Gaussian distribution functions to fit data. The output of GMM is the probability that each data point belongs to each class, from the perspective of prototype clustering, Gaussian mixture clustering describes the prototype by using a probability model, and the cluster partition is determined by the posterior probability corresponding to the prototype, while k-means clustering describes the closeness of the samples in the cluster around the cluster mean vector to a certain extent [13]. In the dataset of this paper, it may be difficult to distinguish the same contributor, and GMM will judge whether the contributor belongs to this category according to the probability that the data points are divided into each cluster. Therefore, in terms of rigor, the classification method using probability to describe data points is much better than the k-means clustering method.

GMM like the k-means algorithm needs to determine the value of super-parameter cluster number \( k \), so we choose the elbow method and silhouette method to determine the value of cluster number \( k \) in this paper. The elbow method is mainly used to express the variance within a cluster, that is, the sparsity of observations within a cluster, while the silhouette method is a comprehensive measure, which is used to express how each individual performs in their cluster. As shown in figure 3, two or four clusters are recommended by the elbow method, while six or eight clusters are recommended by silhouette law. Through analysis, we finally decided to choose four clusters.

![Figure 3. Determination of K value by elbow method and silhouette method.](image)

4. Results and Analysis

In this section, this paper analyzes the results obtained by using the combination of WPCA and GMM as described in the previous section.

Figure 4 shows the principal component loadings relationship between OSM contribution variables and each component (-1 indicates strong negative contribution; +1 indicates strong positive
contribution), and the characteristics of contributors are judged by the loading’s relationship, in this paper, we mainly focus on discovering the characteristics of the behavior created by contributors. Both positive and negative strong contribution loadings reflect the strong correlation between a variable and the principal component.

**Figure 4.** Characteristic contribution load of six main components.

As can be seen from the following figure, PC1 (accounting for 50.6% of the total variance), the contributor characteristic of this kind of description is that the negative value of node modification is extremely large, so it can be said that almost no node modification has been done. There is also very little node creation and update behavior and node improvement, and overall this type of contributor is not very interested in the creation behavior; PC2 (accounting for 17.7% of the total variance), the contributor characteristic of this kind of description is that there are a lot of node improvements, and there are more behaviors for automatic correction of node and improvement of way. However, the creation of node is very few, and the update behavior of node is also few, such contributors are similar to PC1 and are not interested in the creation behavior, PC3 (accounting for 12% of the total variance), the contributor of this kind of description is characterized by a large number of node update behaviors and very few nodes correction. However, the behavior of creating node is less, PC4 (accounting for 7.1% of the total variance), the contributor characteristic of this kind of description is the behavior of modifying a large number of ways. Among them, the behavior of improving the way is a lot. However, there are fewer operations for node, among which the improvement and automatic correction of node are less; PC5 (accounting for 4.2% of the total variance), the contributor characteristic of this kind of description is that a large number of nodes are automatically corrected and more node are created, while the behavior of deleting and correcting node is very few, and the behavior of updating node is also very few. Unlike the previous components, such contributors are more interested in the creation behavior;
PC6 (accounting for 3% of the total variance), which is characterized by a large number of relation modifications. Among them, a lot of relation improvement and more relation automatic correction have been carried out. However, the automatic correction of node takes fewer operations.

In this paper, all OSM contributors in the region are divided into four categories. Table 1 illustrates the characteristics of these four categories respectively.

**Table 1.** Four categories and their related load values on six main components.

|      | PC1     | PC2     | PC3     | PC4     | PC5     | PC6     | n_individuals |
|------|---------|---------|---------|---------|---------|---------|---------------|
| 0    | 979.16  | -8.42   | 40.02   | -5.67   | 26.37   | 5.97    | 8736          |
| 1    | -603909.99 | 303811.03 | 47005.44 | 2576.74 | 19112.37 | 20164.72 | 1             |
| 2    | -10806.85 | -172.87 | -483.94 | 31.72   | -306.43 | -94.85  | 1155          |
| 3    | 900.33   | -6.18   | 32.22   | 2.05    | 20.73   | 7.39    | 5010          |

As can be seen from the above table, different categories of contributors have different characteristics. Group 0 had a total of 8,736 contributors. Of the six component dimensions, PC1 has the highest positive value, followed by PC3, PC5, and PC6, while PC2 and PC4 have negative values close to 0. It can be said that such contributors have hardly made contributions, but have relatively more update behaviors of node, so they are regarded as novices, which is also in line with expectations because the concept of participation inequality put forward by Nielsen and Will Hill [14] shows that most of the data are generated by a few contributors. Group 1, there is only one contributor, and all components except PC1 are very large positive values, so the contribution of this kind of contributor is very large, so this kind of contributor is regarded as the main contributor in this region. Group 2 had a total of 1,155 contributors. Of the six component dimensions, except PC4 is a positive value, all others are negative values, these contributors are keen on the modification and improvement of way, so they are regarded as professional contributors in this region. Group 3 had a total of 5,010 contributors. Of the six component dimensions, except PC2 is negative, all others are positive, and PC1 is the largest. This type of contributor is similar to group 0, but the difference is that this type of contributor contributes a lot to the way, so this type of contributor is regarded as an inexperienced hobbyist in this region.

5. Discussion and Conclusion

In this paper, to classify different contributors according to feature attributes, a weighted principal component analysis method combined with a Gaussian mixture model is proposed. To solve the problem that the meaning of principal components is ambiguous and it is difficult to distinguish sample attributes after dimension reduction, this paper uses WPCA, which can focus on the different characteristics of OSM contributors’ contribution data and the behavior of different characteristics to distinguish different principal components. The output of GMM is the probability that each data point belongs to each class, when it is difficult to distinguish the same contributor, GMM will judge whether the contributor belongs to this class according to the probability that the data points are divided into each cluster. Compared with Jacobs and Mitchell (2020) uses PCA and k-means to classify contributors in a fuzzy way, in this paper, WPCA and GMM are used to classify contributors according to their habit of contributing data.

Through experiments, the contributors are divided into four categories, with Group 1 and Group 2 as the main contributors and professional contributors. At the same time, it further shows that they are the most experienced and professional contributors to the dataset in London, they have contributed a lot of data to the OSM database in this region. Group 0 and Group 4 contributors, as novice and inexperienced contributors, are also important components of OSM, although they make fewer contributions, they enrich the OSM community.

However, the method proposed in this paper also has some problems, that is, the emphasis on feature selection and contribution types of contributors. This paper only considers the weighting value of creating this feature, some interesting phenomena can be observed by weight value to other features. In addition, contributors’ contribution types are also very important, which can further reflect contributors’
contribution habits. Therefore, in the next step, this paper will consider other behavioral characteristics, and consider using the method of information entropy gain to select features.

Generally speaking, checking the behavior characteristics of contributors is a very valuable part of realizing the classification of contributors. Through experiments, it is found that the contributors of OSM can be grouped according to their different behavior characteristics, and finally achieve the classification of contributors.

Acknowledgments
This project was funded by the National Natural Science Foundation of China (41871320).

References
[1] Syaifudin Y W, et al. 2019 The design of road conditions mapping system by utilizing openstreetmap spatial data IOP Conference Series: Materials Science and Engineering 523 (1) 12045.
[2] Zhang L, Pfoser D and Szell M 2019 Using OpenStreetMap point-of-interest data to model urban change-A feasibility study PloS one 14 (2) p. e0212606-e0212606.
[3] Al-Bakri, M and D Fairbairn 2012 Assessing similarity matching for possible integration of feature classifications of geospatial data from official and informal sources International Journal of Geographical Information Science: IJGIS 26 (8) 1437-1456.
[4] Neis P, Zielstra D and Zipf A 2012 The street network evolution of crowdsourced Maps: OpenStreetMap in Germany 2007-2011 Future Internet 4 (1) 1-21.
[5] Heipke C 2010 Crowdsourcing geospatial data ISPRS Journal of Photogrammetry and Remote Sensing 65 (6) 550-557.
[6] Rajaram G and Manjula K R 2019 Exploiting the potential of VGI metadata to develop a data-driven framework for predicting user’s Proficiency in OpenStreetMap context ISPRS International Journal of Geo-Information 8 (11) 492.
[7] Gröchenig S, Rehrl K, Hochmair H H, et al. 2013 A Conceptual Model for Analyzing Contribution Patterns in the Context of VGI (Berlin Heidelberg: Springer).
[8] Budhathoki N R and Haythornthwaite C 2013 Motivation for open collaboration American Behavioral Scientist 57 (5) 548-575.
[9] Jacobs K T and Mitchell S W 2020 OpenStreetMap quality assessment using unsupervised machine learning methods Transactions in GIS 24 (5) 1280-1298.
[10] Shlens J 2014 A tutorial on principal component analysis International Journal of Remote Sensing 51 (2).
[11] Delchambre L 2015 Weighted principal component analysis: A weighted covariance eigendecomposition approach Monthly Notices of the Royal Astronomical Society 446 (4) 3545-3555.
[12] Huang X Y, Li L and Xin Y H 2017 WPCA-LDA: A new data classification method Computer Application Research 34 (006) 1650-1653. (in Chinese)
[13] Zhou Z H 2016 K-Means Machine Learning (Tsinghua University Press) p 202. (in Chinese)
[14] WC Hill, Hollan J D, Wroblewski D A, et al. 1992 Edit wear and read wear Conference on Human Factors in Computing Systems.