International Journal of Geo-Information

MDPI

Article

A Method for Traffic Congestion Clustering Judgment Based on Grey Relational Analysis

Yingya Zhang [1], Ning Ye [1,2,*,†], Ruchuan Wang [3,†] and Reza Malekian [4,*]

1 Department of Computer, Nanjing University of Post and Telecommunications, Nanjing 210003, China; 1014041108@njupt.edu.cn
2 Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing 210003, China
3 Key Lab of Broadband Wireless Communication and Sensor Network Technology of Ministry of Education, Nanjing University of Post and Telecommunications, Nanjing 210003, China; wangrc@njupt.edu.cn
4 Departamento de Ingeniería Informática, Universidad de Santiago de Chile, Av. Ecuador, Santiago 3659, Chile
* Correspondence: yening@njupt.edu.cn (N.Y.); reza.malekian@ieee.org (R.M.); Tel.: +86-138-1389-2478 (N.Y.); +27-12-420-4305 (R.M.); Fax: +86-025-83492470 (N.Y.); +27-12-420-362-5000 (R.M.)
† These authors contributed equally to this work.

Academic Editors: Chi-Hua Chen, Kuen-Rong Lo and Wolfgang Kainz
Received: 17 January 2016; Accepted: 9 May 2016; Published: 18 May 2016

Abstract: Traffic congestion clustering judgment is a fundamental problem in the study of traffic jam warning. However, it is not satisfactory to judge traffic congestion degrees using only vehicle speed. In this paper, we collect traffic flow information with three properties (traffic flow velocity, traffic flow density and traffic volume) of urban trunk roads, which is used to judge the traffic congestion degree. We first define a grey relational clustering model by leveraging grey relational analysis and rough set theory to mine relationships of multidimensional-attribute information. Then, we propose a grey relational membership degree rank clustering algorithm (GMRC) to discriminant clustering priority and further analyze the urban traffic congestion degree. Our experimental results show that the average accuracy of the GMRC algorithm is 24.9% greater than that of the K-means algorithm and 30.8% greater than that of the Fuzzy C-Means (FCM) algorithm. Furthermore, we find that our method can be more conducive to dynamic traffic warnings.

Keywords: urban traffic; grey relational membership degree; traffic congestion judgment

1. Introduction

With the rapid development of urban traffic, urban vehicle surges and the pressure on traffic capacities are increasing sharply. Therefore, traffic problems are becoming serious and bound the development of a city. In China, the conditions of roads and vehicles are quite inconvenient, and traffic congestion has caused substantial social and economic problems. In this case, traffic jams not only waste time, delay work, and reduce efficiency but also cause a substantial waste of fuel, increase the probability of accidents and exacerbate the already serious difficulties facing traffic control and management. Since the 1980s, intelligent transportation systems (ITSs) consisting of integrated computer technology, automatic control technology, communication technology and information processing technology have achieved remarkable results worldwide. In addition, many aspects of ITSs are based on traffic information. Furthermore, traffic information processing has become an important aspect of ITSs [1]. The critical function of an ITS is to manage and control traffic flows and avoid the development of traffic jams. When traffic jams occur, such systems should provide timely and effective solutions and ease traffic pressure. Therefore, clustering and evaluating urban traffic congestion is of

great importance and is thus a prerequisite for correctly inspecting traffic congestion. To determine the road congestion degree, different definitions of traffic congestion are formulated. Rothenberg defined the traffic congestion rank as the number of vehicles on the road exceeding the carrying capacity on the general acceptable road service level [2]. Under such a definition, U.S. authorities divide Level of Service (LOS) into six levels from A to F based on the ratio of the actual vehicle flow (volume) and road capacity: V/C. In Virginia, when V/C is less than 0.77, LOS is at the D level, and the traffic situation is considered to be a high-density but stable traffic flow. When LOS is at the E level, traffic begins to deteriorate and results in a serious traffic jam. A method of assessing the traffic congestion level (rank) is by comparing a certain traffic parameter with a threshold. When the parameter is greater than a certain threshold, a traffic jam is considered to have formed. Specifically, the method can detect whether traffic congestion occurs but cannot represent a comprehensive information evaluation method for traffic congestion. Currently, we collect traffic flow information based on three properties for Nanjing urban trunk roads to comprehensively weigh the level of traffic congestion in the same time period. In addition, we judge which road is allowing smooth traffic flows, which is suffering from a light traffic jam, which is suffering from a traffic jam, and which is suffering from a heavy traffic jam state. Figure 1 represents the Nanjing transport network area in its geographical context. Using the above information, we can provide reference values for traffic management.

At present, the method of judging traffic congestion can be divided into three categories: (1) Direct detection method, such as video detection method. This method needs to install too many cameras and the cost is higher. (2) Indirect detection method, which is mainly according to events' influence on traffic flow, used to detect the event's existence. The method has low cost, simple and easy to operate, but has lower detection rate, and high false alarm rate. (3) Based on theory model, design algorithm to judge traffic congestion. This method is being used, and some mature theories have been applied, such as cluster analysis, grey system theory, and rough set theory. Our work concentrates on clustering traffic flow information based on grey relational analysis to judge traffic congestion situations [3].

Figure 1. Nanjing transport network area in its geographical context.

The paper is organized as follows. First, we give a brief summary of previous related work in Section 2. Next, we introduce how to build the grey relational clustering model in Section 3.

In Section 4, the grey relational membership degree rank clustering algorithm (GMRC) is described. Then, we illustrate experimental results in Section 5. Finally, we conclude the paper in Section 6.

2. Related Work

2.1. Related Theories

Our work particularly involves grey system theory. Grey system theory was proposed by Professor Julong Deng in 1982 and includes many aspects such as grey generation, grey analysis, grey modeling and grey prediction [4]. This theory has been widely used in image processing, network security and logistics management. In addition, grey relational analysis is not only an important aspect of grey theory but also a type of measurement method for the study of the similarity of data. This analysis also has no strict requirement on sample size, and it is usually used for analysis and comparison of the geometric forms of curves described by several points in the space; the closer to the referenced standard array, the higher the relational degree between the referenced standard and the higher its rank. Next, we will briefly introduce rough set theory, which is a mathematical theory method used to address uncertain, imprecise and fuzzy information. This theory has been applied in machine learning, data mining, decision-making analysis, *etc*. In addition, rough set theory is an important branch of uncertainty calculations, which include other theories such as fuzzy sets, neural networks and evidence theory. In this paper, we analyze urban road traffic information using grey relational clustering and combine the results with rough set theory to establish a decision table system. Finally, we judge the degree of urban traffic congestion.

2.2. Clustering Techniques

Clustering analysis is the subject of active research in several fields such as statistics, pattern recognition, machine learning, and data mining. It aims to partition a large number of data into different subsets or groups so that the requirements of homogeneity and heterogeneity are fulfilled. Homogeneity requires that data in the same cluster should be as similar as possible and heterogeneity means that data in different clusters should be as different as possible.

At present, a number of cluster methods have been widely used, among which a weighted adaptive algorithm based on the equilibrium approach is proposed wherein the grey method is introduced to a spectral clustering algorithm [5] to measure the similarity between the data. The experimental results show that the proposed algorithm can effectively overcome the shortcomings of spectral clustering concerning the sensitivity of parameters. Another clustering algorithm based on entropy that can automatically determine the number of clusters based on the distribution characteristics of the data sample and reduce the user's participation was proposed in [6]. The result is more objective, and the algorithm can find large and small clusters of any shape. The disadvantages of the algorithm are the selection of the initial points and the effects of noise and outliers. However, these shortcomings can be overcome by screening and addressing the original data noise and outliers, excluding false data and improving the reliability and separation of the data to minimize the influence of noise and outliers. The two methods inspire our work such that we consider using an approach to comprehensively evaluate data; therefore, we use grey relational theory, which can effectively process multidimensional attribute data. However, those traditional algorithms are mostly a simple clustering of similar data and do not consider what data attributes have what indicator characteristics. Moreover, clustering results cannot reflect the rank of data that present greater clustering.

2.3. K-Means Algorithm and FCM Algorithm

Currently, the K-means algorithm and Fuzzy C-Means (FCM) algorithm are commonly used to cluster data. We consider the two algorithms comparing with our proposed algorithm in our work. First, our work will provide a brief introduction of the K-means algorithm. The K-means algorithm is a classical clustering algorithm that is widely used in different subject areas. In addition, various

improved algorithms have evolved based on the K-means algorithm. The K-means algorithm, however, is an NP-hard optimization problem (Generally speaking, problems that will cost polynomial time to solve and are easy to address are commonly regarded as P problems. Problems that cost super polynomial time to solve are considered as difficult problems and are known as NP-hard problems), which means that many problems cannot obtain global optimal results [7]. The Euclidean distance is typically used as the criterion function of the K-means algorithm, where the distance between two data points with different units is sometimes calculated. The Euclidean distance is the real distance between two points in an m-dimensional space and is mainly determined by the heavily fluctuant elements. Slightly fluctuant elements are often neglected, and the phenomenon is more obvious with increasing ratio of the difference between corresponding elements.

FCM algorithm is a type of fuzzy clustering algorithm based on the objective function. First, we introduce the concept of fuzziness: fuzziness is uncertainty. In addition, we usually say that an object is what is certain. However, the uncertainty indicates the similarity between two objects [8,9]. For example, we regard twenty years of age as a standard of judging whether an individual is young or not young. Therefore, a 21-year-old person is not young according to this division of certainty; however, 21 years of age is very young in our opinion. Thus, at this moment, we can vaguely think of the possibility of a 21-year-old person belonging to young as 90% and that belonging to not young as 10%. Here, 90% and 10% are not probabilities; rather, they are the degree of similarity. Although the FCM algorithm can effectively perform clustering, it does not differentiate the rank to which the clustering belongs.

In view of the above problems, we mine traffic flow information relations with different attributes via grey relational clustering. In addition, we propose the GMRC algorithm to judge the clustering priority. Simultaneously, we combine the K-means and FCM clustering algorithms and contrast the results with those of the GMRC algorithm to evaluate the performance of our algorithm.

3. Grey Relational Clustering Model

In this paper, we suppose that the traffic congestion state of roads is divided into four ranks (not four clusters), namely, smooth, light jam, jam and heavy jam, which is our precondition, where smooth characterizes the best condition, belonging to the first rank, and heavy jam characterizes the worst condition, belonging to the fourth rank. According to the description of the different definitions of traffic congestion, traffic congestion is not only related to certain parameters [10] (such as traffic volume and traffic speed) but also includes many factors. Therefore, a single parameter used to describe traffic congestion states is insufficient: when the vehicle's speed is zero, it may be blocked by too many vehicles on the road that cannot move or it may also be smooth, with no vehicles driving on the road. Therefore, a light traffic flow can match two states: heavy jam and smooth. In addition, low density may characterize traffic that is smooth and may also include more trucks and other large vehicles on the road. However, comprehensive analysis of the three variables of traffic flows (traffic flow velocity, traffic flow density and traffic volume) can reflect the real situation concerning traffic jams. Our purpose is to evaluate the traffic jam degree of different roads based on a multiple-attributes index. Thus, we introduce three variables of traffic flows to evaluate the degree of traffic jams. However, to consider addressing data of multidimensional mixed attributes and obtain data clustering levels, we use grey relational analysis theory.

Traffic flows are characterized by three properties [11]: traffic flow velocity, traffic flow density and traffic volume. Traffic flow velocity indicates the average speed of vehicles on the road, in units of km/h. The traffic flow density, namely, the density of vehicles, indicates the number of vehicles per unit length that the road contains. The traffic volume is defined as the number of vehicles traveling through a certain road section in unit time. This paper uses these three properties of traffic flows to judge the traffic congestion state.

Definition 1. *Assume the analysis domain data set* $X = \{X_i | X_i = (X_{i1}, X_{i2}, X_{i3})\} \quad i \in N$ *as the comparative object set, where* X_i *represents the ith data object, each of which has three attributes: traffic flow velocity, traffic flow density, and traffic volume.*

Definition 2. *Suppose a reference standard array set as* $Y = \{Y_i | Y_i = (Y_{i1}, Y_{i2}, Y_{i3})\}$ $(i = 1, 2, \ldots, p)$ *(referenced object set). We regard the traffic flow information (when the road is in the smooth state or vehicles are traveling at the free stream velocity) as a reference standard array set.*

Each group of a reference sequence can obtain a clustering result when combined with a comparative object set by the grey relational clustering method; therefore, p groups of referenced arrays will produce p clustering results. In addition, the group of referenced standards extracted from the data is used for clustering when combined with comparative object sets, which can produce a clustering result. Thus, over the whole process, we can obtain $p + 1$ clustering results.

Definition 3. *The evaluation information M is regarded as the output information of the grey relational clustering system S, and F is defined as the evaluation function. Then, the relationship between S and M is denoted as*

$$\begin{cases} S = ((X, Y), G) \\ F : X \times Y \rightarrow M \end{cases} \quad (1)$$

where G represents the grey relational similar matrix. The evaluation function F of the above model leverages the decision table system of rough set theory to weigh the degree of contribution of the cluster members to the clustering results. F is also called the grey relational membership function, whose inputs are X and Y and output is M, which reflects the similarity between elements inside a class and the similarity between classes.

In this paper, we propose the GMRC algorithm oriented to multidimensional attribute data to judge clustering rank priority. The method's procedure includes pre-processing data, transforming the data into matrix form, setting the threshold, and filtering and deleting abnormal data objects. Next, we cluster analysis domain data objects via grey relational clustering analysis and then apply weighted decision analysis from rough set theory. Finally, the clustering results are calculated using probability theory.

4. GMRC Algorithm

The architecture of the GMRC algorithm consists of the following phases (Figure 2):

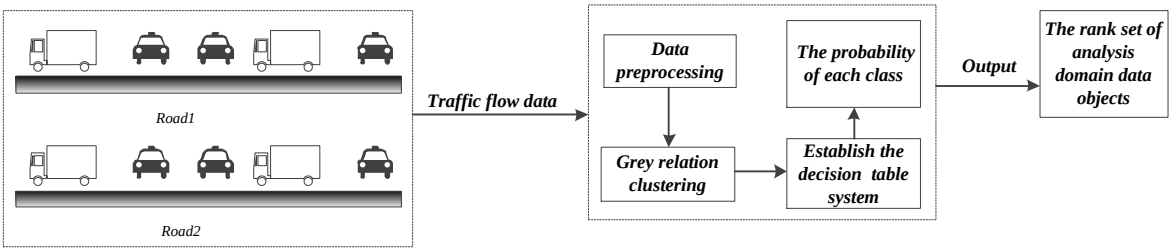

Figure 2. The architecture of the grey relational membership degree rank clustering algorithm (GMRC).

According to the index feature of the three properties of the traffic flow data, the optimal referenced standard is extracted from the data set of the analysis domain. Then, the optimal referenced standard and referenced array set are used to calculate the grey relational degree in combination with the dataset of the analysis domain. Subsequently, the grey relational matrix can be obtained, and the cluster members can be calculated. Next, we use rough set theory, where cluster members are applied, and build the decision information table to complete the fusion for the weighting of the cluster members, according to which we need to calculate the rank of each data object using probability theory. We regard the first category as the highest priority, which means that the data object is closer to the referenced standard and simultaneously that the object is better, corresponding to the lightest traffic congestion degree.

4.1. Grey Relational Clustering Steps

4.1.1. Extracting Optimal Referenced Standard according to the Characteristics of Traffic Flow Data

Because the problem of clustering analysis is solved based on a given indicator system, it is very important to choose the appropriate indicator to achieve reasonable and appropriate clustering. There are three properties for each X_i from a data source, and based on the indicator attribute of the data object, the optimal reference standard X_0 can be extracted from the data source. Specific instructions are as follows.

We know that the three properties units are different; therefore, we first need to convert the data to the same format. "↑" indicates that the property value is greater and thus better; "↓" indicates the opposition. *(a, b)*, where *a* and *b* are numbers, indicates that, if the property value of a data object is in this interval, the value will be better.

Extract the optimal referenced standard X_0 according to the characteristics of the three properties, and its expression is $X_0 = \{X_{01}, X_{02}, X_{03}\}$, which is described as follows:

Considering the characteristic of the traffic flow velocity as "↑", we use

$$X_{0j} = \max X_{ij} \ (i \in N, j = 1) \tag{2}$$

Considering the characteristic of the traffic flow density as "↓", we use

$$X_{0j} = \min X_{ij} \ (i \in N, j = 2) \tag{3}$$

Considering the characteristic of the traffic flow volume as "↓", we use

$$X_{0j} = \left\{ X_{ij} | \min \left| X_{ij} - \frac{(a+b)}{2} \right| \ (i \in N, j = 3) \right\} \tag{4}$$

4.1.2. Data Normalization Processing

Because there are different types of data, the units are also different. According to the characteristics of the properties, the data of the analysis domain are processed using different measures, and the data are compressed to (0, 1). The processing is as follows:

For the traffic flow velocity of the whole data set, we use

$$X_i(j) = \left| \frac{X_{ij} - X_{j\,min}}{X_{j\,max} - X_{j\,min}} \right| \ (i = 0, 1, \ldots, n, j = 1) \tag{5}$$

For the traffic flow density of the whole data, we use

$$X_i(j) = \left| \frac{X_{ij} - X_{j\,max}}{X_{j\,max} - X_{j\,min}} \right| \ (i = 0, 1, \ldots, n, j = 2) \tag{6}$$

For the traffic volume of the whole data set, we use

$$X_i(j) = \begin{cases} 1 - \left| \frac{X_{ij} - X_{0j}}{X_{0j}} \right|, & \left| \frac{X_{ij} - X_{0j}}{X_{0j}} \right| \leqslant 1 \ and \ i = 0, 1, \ldots, n, j = 3 \\ 0, & \left| \frac{X_{ij} - X_{0j}}{X_{0j}} \right| > 1 \ and \ i = 0, 1, \ldots, n, j = 3 \end{cases} \tag{7}$$

X_{ij} is the original matrix element in the above Equations (5)-(7) and is normalized to $X_i(j)$, where $X_{j\,max}$ represents the maximum of the *jth* column and $X_{j\,min}$ represents the minimum of the *jth* column, and anything inside of braces is the limiting condition. After normalization, we obtain $(p + 1)$ matrices, namely, $A_0, \ A_1, \ldots, \ A_p$ where A_0 can be acquired by normalizing the optimal referenced standard X_0 and the comparative object set. Meanwhile, $A_1, \ A_2, \ldots, \ A_p$ can be acquired by normalizing the referenced array set and the comparative object set, where A_0 and A_p are defined as

$$A_0 = \begin{bmatrix} X_1(1) & X_1(2) & X_1(3) \\ X_2(1) & X_2(2) & X_2(3) \\ \cdots & \cdots & \cdots \\ X_n(1) & X_n(2) & X_n(3) \\ X_0(1) & X_0(2) & X_0(3) \end{bmatrix} \quad A_p = \begin{bmatrix} X_1(1) & X_1(2) & X_1(3) \\ X_2(1) & X_2(2) & X_2(3) \\ \cdots & \cdots & \cdots \\ X_n(1) & X_n(2) & X_n(3) \\ Y_p(1) & Y_p(2) & Y_p(3) \end{bmatrix} \tag{8}$$

The last row of the matrix A_0 is normalized to the optimal referenced standard sequence, and the last row of the matrix A_p is the *pth* normalized referenced standard array.

4.1.3. Calculating Grey Relational Degree and Generating Grey Relational Similar Matrix

(1) The grey relational degree reflects the high degree between two comparative objects. For example, we focus on calculating the grey relational degree of the matrix A_0. The formulas are as follows (Equations (9) and (10)):

$$\gamma_{0i(k)} = \frac{\min\limits_i \min\limits_k |X_0(k) - X_i(k)| + \sigma \max\limits_i \max\limits_k |X_0(k) - X_i(k)|}{|X_0(k) - X_i(k)| + \sigma \max\limits_i \max\limits_k |X_0(k) - X_i(k)|} \quad (k = 1, 2, 3) \tag{9}$$

$$\gamma_{0i} = \frac{1}{m} \sum_{k=1}^{m} \gamma_{0i}(k) \tag{10}$$

where σ is the resolution coefficient, which has a range of 0 to 1. Generally, we assume that σ is 0.5. $\gamma_{oi}(k)$ is the correlation coefficient between X_i and X_0 at the *kth* point. The grey relational degree is expressed by $\gamma_{oi}(k)$, where X_0 is regarded as the referenced sequence and X_i is regarded as the comparative sequence. Similarly, we can obtain γ_{ij} when X_i is regarded as the referenced sequence and X_j is regarded as the comparative sequence. Then, X_1, $X_2, \ldots,$ X_n, X_0 are regarded as the referenced sequence; meanwhile, the $(n + 1)$ sequences are regarded as comparative sequences (the $(n + 1)$ sequences are not only referenced sequences but also comparative sequences). Finally, we calculate the grey relational degree matrix $\Gamma^0_{(n+1) \times (n+1)}$ according to the grey relational analysis, where $\Gamma^0_{(n+1) \times (n+1)}$ is obtained based on A_0 and consists of any γ_{ij} $(i = 1, 2, \ldots, n, n+1, j = 1, 2, 3)$. Similarly, we can calculate the grey relational degree matrices Γ^1, $\Gamma^2, \ldots, \Gamma^p$ based on A_0, $A_1, \ldots,$ A_p.

(2) Calculate grey relational similar matrix G

We calculate the similarity elements $g_{ij} = (\gamma_{ij} + \gamma_{ji})/2$ in G. G_0, for example, is the grey relational similar matrix obtained when X_0 is regarded as the referenced sequence, and the $(n + 1)th$ row grey relational similarity elements of G_0 are calculated when X_0 is regarded as the optimal referenced sequence. Similarly, we can obtain $p + 1$ grey relational similar matrices: $G_0, G_1, \ldots, G_p$.

4.1.4. Grey Relational Degree Clustering Analysis

Based on G, *i.e.*, the grey relational similar matrix, we construct the maximal relational tree, which consists of the values of the last row elements ordered in G. This is shown in Figure 3.

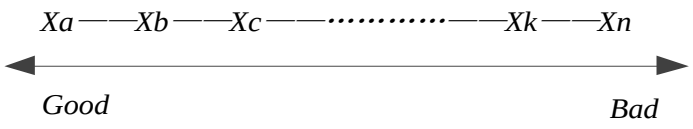

Figure 3. Maximal relational tree.

Based on G_0, we can obtain g_{ij}, which is also called the closeness degree between X_i and X_j. The greater g_{ij} is, the closer X_i and X_j are; in contrast, X_i and X_j are further away for decreased

g_{ij}. Based on Figure 3, the maximal relational tree with closeness degrees is generated, as shown in Figure 4, where a_i represents the closeness degree between X_i and X_j.

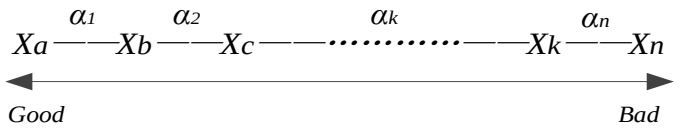

Figure 4. Maximal relational tree with closeness degrees.

Based on Figure 4, we set the isolated point coefficient λ, which is in the interval (0, 1). We cut the tree when the closeness degree is less than λ and when adjacent branches exhibit substantial differences. Therefore, the disconnected tree is used to make its connected branches form k levels of clusters along the horizontal. We consider the closest branch as the first level and the loosest branch as the *kth* level (If $k \geqslant 4$, we also classify the fifth branch, the sixth branch, the seventh branch, and even the *kth* branch as the fourth rank. This is to say that smooth corresponds to the first rank and that heavy jam corresponds to the fourth rank). Thus, in this way, we can obtain a cluster member of G. Therefore, we use Γ^0, $\Gamma^1, \ldots,$ Γ^p to compute $p + 1$ grey relational similarity matrices G_0, G_1, $\ldots$, G_p, from which we can find the closeness relationship among each object. Therefore, we can obtain $p + 1$ clustering results, namely, cluster members.

4.2. Establishing Evaluation Function for Grey Relational Clustering System

According to the above initial clustering results, we use rough set theory to establish the decision table system that is applied to weight the contribution of cluster members to the clustering results and give weights to the cluster members.

4.2.1. Describing How to Establish the Decision Table System

First, the optimal referenced standard and the referenced standard set are combined with the comparative object set through the grey relational clustering method to construct the decision table system $F = \langle U, C, D, V, f \rangle$,where $U = \{X_1, X_2, \ldots, X_n\}$ represents analysis domain data; $C = \{c_1, c_2, \ldots, c_p\}$ are conditional attributes and are cluster members formed by the referenced standard set; $D = \{d\}$, as the decisional attribute, is the cluster member obtained using the optimal referenced standard; $V = V_C \cup V_D, V_C = \cup V_{c_h}$, $c \in C$ are the range of the set of traffic flow properties, where V_{c_h} represents the level in cluster member c_h ($h = 1, 2, \ldots, p$); f represents the evaluation function, $f : U \times (C \cup D) \rightarrow V$; and $f(X_i, c_h) \in V_{c_h}$ represents the level of X_i in cluster member c_h.

4.2.2. Calculating Information Entropy

In the decision table system, the information entropy weight $I(c_h, D)$ indicates how important the cluster member c_h (conditional information) is for result D (decisional information) when the optimal referenced standard is chosen to calculate the cluster member. According to the information entropy of rough set theory, $I(c_h, D)$ is described as follows:

$$I(c, D) = H(D) - H(D|\{c\}) \tag{11}$$

$$H(c) = -\sum_{i=1}^{k} P(RC_i) \log(P(RC_i)) \tag{12}$$

$$H(D|\{c\}) = -\sum_{i=1}^{k} P(RD_i|RC_i) \log(P(RD_i|RC_i)) \tag{13}$$

$$P(RC_i) = \frac{|RC_i|}{|X|} \tag{14}$$

$$P\left(RD_i|RC_i\right) = \frac{|RD_i \cap RC_i|}{|RC_i|} \tag{15}$$

where $i = 1, 2, \ldots, k$ (k is the number of clusters), RC_i represents the *ith* divided cluster of cluster member c, $|RC_i|$ represents the number of elements in the *ith* cluster, and $|X| = n$. A larger conditional attribute c is more important to the decisional information D. In addition, $H\left(c\right)$ and $H(D|\{c\})$ are determined by the conditional information entropy of rough set theory. Thus, the relative weight of each cluster member can be determined, and a more important cluster member corresponds to a greater weight.

4.3. Calculating the Level of Clustering Membership of Data Objects

Step 1: Calculate the importance of the attribute information entropy $E_h = I\left(c_h, D\right)$ for each cluster member c_h in the decision system, where $h = 1, 2, \ldots, p$.

Step 2: Set the relative weight of each cluster member:

$$\omega_h = E_h / \sum_{h=1}^{p} E_h \tag{16}$$

Step 3: Use probability theory to calculate the probability of each data object emerging in every clustering based on the relative weights to choose the level whereby the probability is maximized. Furthermore, obtain the final clustering results. In addition, data object X_i belonging to the *jth* level ($j = 1, 2, \ldots, k$) is defined as

$$P\left(X_i^j\right) = \sum_{\substack{h=1 \\ M_{ik}=j}}^{p} w_h \tag{17}$$

where M_{ik} represents the level of data object X_i in cluster member c_h, which has been computed in grey relational clustering. Thus, the grey relational membership degree level of X_i can be expressed as:

$$Level\left(X_i\right) = \{j|\underset{j=1\rightarrow k}{max}(P(X_i^j))\} \tag{18}$$

The final result is $C = \left\{C^1, C^2, \ldots, C^k\right\}$, where C^k includes all data objects whose grey relational membership degree level is the *kth* level:

$$C^k = \{X_i|Level\left(X_i\right) = k, X_i \in X\} \tag{19}$$

4.4. GMRC Algorithm Detail Description

In this paper, we study the problem of multidimensional-attribute information clustering for traffic flow and propose the GMRC algorithm. First, we transform the dataset into matrix form, extract the optimal referenced standard from the dataset, and then perform the normalized processing to eliminate the effects of different units. Furthermore, we obtain the preliminary clustering results according to grey relational theory analysis. Finally, we build a decision table system to calculate the relative weight for each cluster member. The algorithm is described as follows.

Input: Analysis domain data set $X = \{X_i|X_i = (X_{i1}, X_{i2}, X_{i3})\}\ i \in N$,
Referenced array set $Y = \{Y_i|Y_i = (Y_{i1}, Y_{i2}, Y_{i3})\}\ (i = 1, 2, \ldots, p)$.

Output: The level (rank) set of analysis domain data objects

Algorithm 1: GMRC (X,Y)

```
1  Level = null; Weight = null; Member = null; Entropy = null;
   //Initialize sets Level, Weight; Matrix Member, Entropy
2  Data_ Preprocessing (X,Y);              // Data pre-processing
3  X0 = ExtractOptimal (X);                // Extract the optimal referenced standard
4  S = Normalization (X,Y);                // Normalization processing
5  T = MaxRelTrees (S)                     // Construct the maximum relational trees
6  T' = ClosenessTrees (T)                 // Construct the maximum relational tree with closeness degree
7  Member0 = GreyCluster (X,X0);           // Regard X0 as referenced standard to compute cluster member
8  Foreach (Yi in Y)
9  Memberi = GreyCluster (X,Yi);           // Regard Yi as referenced standard to compute cluster member
10 End
11 F = DecisionSystem (Members);           // Establish the decision table system F
12 Entropy = CalculateEntropy (F);         // Calculate the information entropy for each cluster member
13 Weight = CalculateWeight (Entropy);     // Calculate relative weights for each cluster member
14 CalculateLevel (X);                     // Calculate membership degree level of Xi in X
```

The above steps of the algorithm are described as follows:

Step 1: Initialize the parameters, pre-process traffic flow data, set the threshold to filter and delete abnormal data objects (lines 1 and 2).

Step 2: According to the features of the three properties of the traffic flow data, extract the optimal referenced standard from the analysis domain data (line 3).

Step 3: Normalize the analysis domain data set in combination with the referenced standard sequences (line 4).

Step 4: Compute the grey relational degree of the corresponding matrix; further, determine the grey relational similarity matrices and then construct the maximum relational trees based on the $(n + 1)th$ row elements of those matrices (line 5).

Step 5: Based on Step 4, construct the maximum relational tree with closeness degrees (line 6).

Step 6: Compare the closeness degree between data objects, cut off the tree when the closeness degree is less than λ and adjacent branches exhibit large differences, and then obtain k levels of clustering results. Similarly, we obtain in total $p + 1$ cluster members from the $p + 1$ referenced arrays (lines 7–10).

Step 7: Establish the decision table system based on p cluster members as conditional attributes obtained from the referenced standard array set. In addition, the only cluster member obtained from the optimal referenced standard (line 11) is regarded as the decisional attribute.

Step 8: Compute the information entropy of cluster members for decision making, which is used to weigh the contribution of each cluster member to the clustering results (line 12).

Step 9: Calculate the weight of each cluster member (line 13).

Step 10: Calculate the probability of each data object emerging in every clustering; then, choose the level when the probability is maximized. Furthermore, obtain the final clustering results (line 14).

4.5. Grey Relational Membership Function

The performance of the traditional clustering results depends on the distance between elements inside classes and the distances between classes. Shorter distances between elements inside a class indicate better classes; conversely, longer distances between classes indicate better classes [12]. Our purpose for clustering is to obtain the membership degree rank of classes and to judge the rank of data objects. Thus, we construct a membership function reflecting the similarity between elements inside a class and the similarity between classes based on the grey relational similarity degree. Assume that $\gamma_{X,Y}$ is denoted as the grey relational similarity degree between object X and object

Y, and $C = \left\{C^1, C^2, \ldots, C^k\right\}$ represents divided clusters. S_C, which is used to weigh the division C, is defined as

$$S_c = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{\left|C^i\right| \times \left|C^i\right|}\sum_{X,Y\in C^i}\gamma_{XY} + \frac{1}{k\,(k-1)}\sum_{i=1}^{k}\Big(\sum_{i=1,l\neq i}^{k}\frac{1}{\left|C^i\right| \times \left|C^l\right|}\sum_{X\in C_i, Y\in C^l}\gamma_{XY}\Big) \tag{20}$$

5. Experimental Results and Analysis

In this paper, we collected traffic data for Nanjing trunk roads, which connect the main commercial areas and which include high traffic volumes and high-density residential areas. These roads need to be cleared in a timely manner before traffic jams can occur. Generally speaking, people's daily routines are very regular [13]: go to work in the morning and go home at dusk. However, this can quickly lead to traffic congestion during rush hours. A previous study noted that a single parameter, such as velocity, used to describe traffic congestion states is insufficient. To obtain detailed knowledge for judging traffic congestion, comprehensive analysis of the three variables of traffic flows for determining the traffic flow state can be used to reflect the real conditions of the road for predicting traffic jams [14]. To test our algorithm, we experimentally collected traffic flow data from 50 monitoring points along Nanjing's trunk roads during the time periods of approximately 7:00–9:30 a.m. and 4:30–7:00 p.m., which correspond to the rush hours. In addition, we chose 30 drivers with more than five years of driving experience as testers and watched their vehicle driving videos to obtain their evaluation of the traffic flows' four states (smooth, light jam, jam, and heavy jam) [15]. Then, we evaluated the clustering results to validate the accuracy of our algorithm compared with other clustering methods such as the K-means algorithm and the FCM algorithm.

In this experiment, we assumed that the resolution coefficient σ is 0.5 and that the number of levels is 4. Because the algorithm is stochastic in nature, the average results of 20 tests for each algorithm on each dataset are used as the experimental results. Table 1 shows the corresponding information entropy of the four cluster members obtained by the primary clustering data samples, namely, the relative weights of the cluster members. To verify the performance of our algorithm, six sample points were randomly selected from the dataset, as shown in Table 2. In Table 2, we find that the clustering of the traffic state is most closely related to traffic flow velocity; however, the velocity cannot fully determine the traffic state. For example, the traffic flow speed in the 459th group is 40.3 km/h, and the state is a light jam. However, the traffic flow speed in the 812th group is 45.5 km/h, which is greater than 40.3 km/h, but the state is a jammed state. This phenomenon occurs mainly because the traffic volume of the latter is greater than that of the former. The three properties of the traffic flow should be comprehensively considered rather than only the velocity.

Table 1. Relative weights of cluster members.

Cluster Members	Relative Weights
C1	0.3746
C2	0.2644
C3	0.2128
C4	0.2481

Table 2. Random samples.

Samples	Traffic Flow Velocity	Traffic Flow Density	Traffic Volume	Results
15	66.0	58.2	16.2	Smooth
35	35.2	70.3	25.6	Jam
106	50.8	66.9	23.9	Smooth
459	40.3	68.3	31.2	Light Jam
689	10.8	81.9	41.7	Heavy Jam
812	45.5	63.6	58.9	Jam

5.1. Complexity Analysis

We suppose k levels, n data objects, m dimensions, and p referenced standard sequences. The complexity of the GMRC algorithm is described in the following.

5.1.1. Time Complexity Analysis

The time for constructing the initial matrix is $O(m \times n)$. Extracting the optimal referenced standard requires access to all elements in the initial matrix, and the time complexity is also $O(m \times n)$. The time complexity of the matrix normalization is $O(p \times n \times m)$. The grey relational degree calculation requires access to all the elements in n grey relational similarity degree matrices; therefore, its time complexity is approximately $O(n \times m \times p \times n)$. The time complexity of sorting the $(n+1)th$ row grey relational degree elements for p grey relational similarity degree matrices is $O\left(p \times n^2\right)$. Calculating the information entropy for the cluster members needs $O(p \times k)$. The time needed to calculate the clustering membership degree analysis domain data is $O(n \times p \times k)$. According to the above analysis, the average time complexity of the GMRC algorithm is $O\left(k \times m \times p \times n^2\right)$.

5.1.2. Space Complexity Analysis

The complexity of analyzing the domain initial matrix is $O(m \times n)$ in space, and the complexity of analyzing the similarity degree matrix with space complexity is $O\left(p \times n^2\right)$. In addition, the space complexity of determining the grey relational membership degree in the GMRC algorithm is $O(k \times p)$. Therefore, the total space complexity of the algorithm is $O\left(k \times p \times n^2\right)$.

5.2. Impact of Isolated Point Coefficient λ on the Clustering Results

From Figure 5, we can find that, based on different numbers of data samples, the grey relational membership function value gradually increases with increasing λ. This is because with increasing λ, the grey relational similarity degree between the elements inside classes increases, and this accounts for the dominant position. Clearly, when λ is between 0.84 and 0.87, the function value is maximized. Then, as λ continues to gradually increase, the function value decreases. This is because the grey relational similarity degree between classes occupies the dominant position. From the perspective of the grey relational similarity degree between data objects, we analyze the closeness degree inside classes and among classes and fully utilize the multi-dimensional information feature and the overall change in the three properties to better describe the closeness degree between data objects. Therefore, in the next experiment, the range of λ is set as (0.84, 0.87).

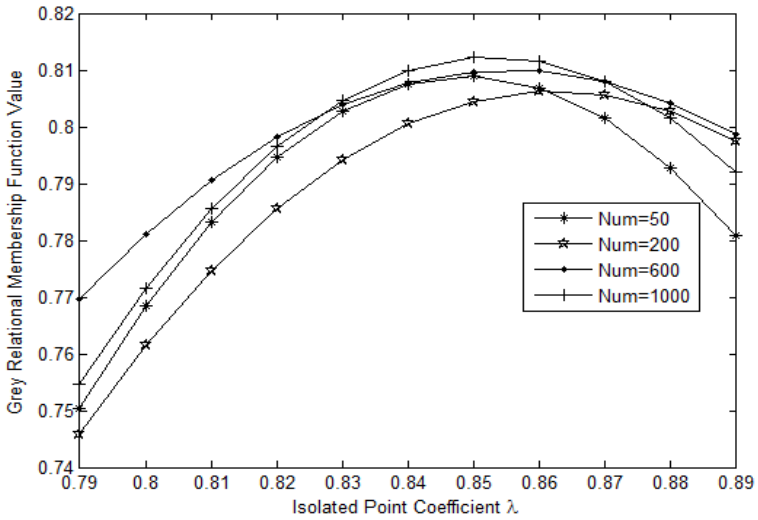

Figure 5. Grey relational membership function as a function of λ.

5.3. Comparison with Other Algorithms

Figure 6 illustrates the accuracy rate of the GMRC, K-means and FCM clustering algorithms in each class. We find that the accuracy of GMRC in each class and under different λ is higher than that of the K-means and Fuzzy algorithms. This is because the GMRC algorithm uses the data attribute index feature to compute the grey relational similarity degree and takes the quality of cluster members into consideration. Thus, the algorithm effectively improves internal class similarity to achieve ranked clustering.

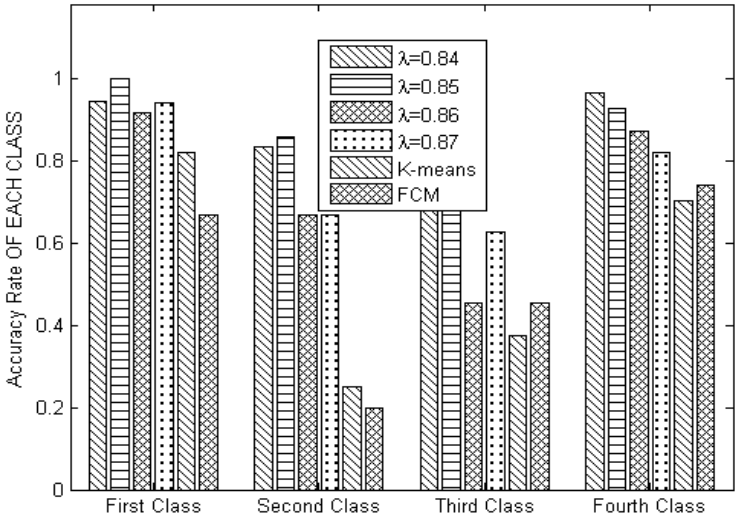

Figure 6. Comparison of accuracy rate of each class among GMRC, K-means and FCM algorithms.

Figure 7 shows the average accuracy of our GMRC algorithm under different λ and that of the K-means and FCM algorithms. From Figures 6 and 7 we can conclude that the average accuracy of the GMRC algorithm is higher than that of the K-means and FCM algorithms. In addition, the average accuracy of the GMRC algorithm is 24.9% higher than that of the K-means algorithm and 30.8% higher than that of the FCM algorithm. In addition, our new algorithm exhibits better stability. Because our algorithm does not need to randomly choose the initial center point, as in the K-means algorithm, the stability of the algorithm is not affected by the stochastic nature of the algorithm.

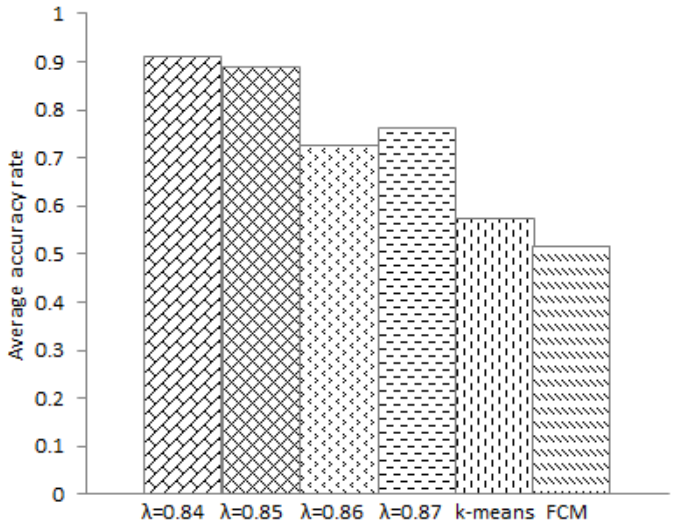

Figure 7. Comparison of average accuracy rate among the GMRC algorithm with different λ and the K-means and FCM algorithms.

6. Conclusions

Judging traffic congestion states is the premise and basis for dynamic traffic congestion warning, traffic guidance, actively avoiding traffic congestion and ensuring smooth roads. However, traffic jams are usually judged by experience. This paper collects traffic flow data and provides a more effective judgment method. We introduce both grey relational analysis and rough set theory to the GMRC algorithm and weigh the membership degree of data object clustering using comprehensive information about the data. In this process, we construct the maximum relational tree with closeness degree and compare the closeness degree between data objects. We cut off the tree when the closeness degree is less than λ and when adjacent branches exhibit large differences. Consequently, we obtain $p + 1$ cluster members. Next, we establish a decision table system based on p cluster members as conditional attributes obtained from referenced standard array sets. Then, we calculate the probability of each data object emerging in every clustering, choose the rank when the probability is maximized, and finally obtain the final clustering results. Thus, our algorithm fills the gaps present in the literature whereby the K-means and FCM algorithms cannot differentiate which rank a clustering belongs to. The experimental results show that the proposed algorithm, which takes the characteristics of the multidimensional data object attributes into consideration, is a superior algorithm. Next, we plan on applying grey relational similarity to other algorithms and to consider reducing the computational complexity of the algorithm.

Acknowledgments: This research was performed in cooperation with the Institution. The research is support by the National Natural Science Foundation of China (No. 61572260, No. 61373017, and No. 61572261), the Peak of Six Major Talent in Jiangsu Province (No. 2010DZXX026), the China Postdoctoral Science Foundation (No. 2014M560440), the Jiangsu Planned Projects for Postdoctoral Research Funds (No. 1302055C), the Scientific & Technological Support Project of Jiangsu Province (No. BE2015702), the Jiangsu Provincial Research Scheme of Natural Science for Higher Education Institutions (No. 12KJB520009), and the Science & Technology Innovation Fund for Higher Education Institutions of Jiangsu Province (No. CXZZ11-0405). The authors are grateful to the anonymous referee for a careful review of the details and for their helpful comments, which improved this paper.

Author Contributions: All the authors conceived of and designed the study. Furthermore, Yingya Zhang designed the GMRC algorithms presented in this paper and produced the results. Ning Ye provided guidance on modeling, Ruchuan Wang provided guidance on theory and publication fee, and Malekian provided guidance on simulation.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Courtney, R.L. A broad view of its standards in the U.S. In Proceedings of the IEEE Conference on Intelligent Transportation Systems, Boston, MA, USA, 9–12 November 1997; pp. 529–536.
2. Arnold, E.D., Jr. Congestion on Virginia's Urban Highways. Available online: http://ntl.bts.gov/DOCS/arnold.html (accessed on 17 April 1998).
3. Mok, P.Y.; Huang, H.Q.; Kwok, Y.L.; Au, J.S. A robust adaptive clustering analysis method for automatic identification of clusters. *Pattern Recognit.* **2012**, *45*, 3017–3033. [CrossRef]
4. Yu, B.; Zhou, Z.; Xie, M. New grey comprehensive correlation degree model and its application. *Technol. Econ.* **2013**, *32*, 108–114.
5. Hu, F.; Xia, G.-S.; Wang, Z.; Huang, X.; Zhang, L.; Sun, H. Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 2015–2030.
6. Li, K.; Li, P. Fuzzy Clustering with generalized entropy based on neural network. In *Unifying Electrical Engineering and Electronics Engineering*; Springer: New York, NY, USA, 2014; pp. 2085–2091.
7. Zhang H, R.; Zhang, F. The traditional K-means clustering algorithm research and improvement. *J. Xianyang Norm. Univ.* **2010**, *25*, 138–144.
8. Zheng, Y.; Jeon, B.; Xu, D.; Wu, Q.M.J.; Zhang, H. Image segmentation by generalized hierarchical fuzzy C-means algorithm. *J. Intell. Fuzzy Syst.* **2015**, *28*, 961–973.
9. Li, K.; Cui, L. A kernel fuzzy clustering algorithm with generalized entropy based on weighted sample. *Int. J. Adv. Comput. Res.* **2014**, *4*, 596–600.

10. Wen, H.; Sun, J.; Zhang, X. Study on traffic congestion patterns of large city in China taking Beijing as an example. *Procedia-Soc. Behav. Sci.* **2014**, *138*, 482–491. [CrossRef]
11. Stefanello, F.; Buriol, L.S.; Hirsch, M.J.; Pardalos, P.M.; Querido, T.; Resende, M.G.C.; Ritt, M. On the minimization of traffic congestion in road networks with tolls. *Ann. Oper. Res.* **2015**. [CrossRef]
12. Guan, X.; Sun X, W.; He, Y. A novel feature association algorithm based on grey correlation grade and distance measure. *Radar Sci. Technol.* **2013**, *4*, 363–367, 374.
13. He, H.; Tang, Q.; Liu, Z. Adaptive correction forecasting approach for urban traffic flow based on fuzzy C-Mean clustering and advanced neural network. *J. Appl. Math.* **2013**, *2013*, 633–654.
14. Lu, X.; Song, Z.; Xu, Z.; Sun, W. Urban traffic congestion detection based on clustering analysis of real-time traffic data. *J. Geo-Inf. Sci.* **2012**, *14*, 775–780. [CrossRef]
15. Elefteriadou, L.; Srinivasan, S.; Steiner, R.L.; Tice, P.C.; Lim, K. Expanded transportation performance measures to supplement level of service (LOS) for growth management and transportation impact analysis. *Congest. Manag. Syst.* **2012**, *19*, 977–991.