Multivalued function recognition based on spectral clustering

Zongchao Huang¹, Zhaogong Zhang²*, and Guanwen Yu³

¹,²,³ School of Computer Science and Technology, Heilongjiang University, Heilongjiang Province, China

*Corresponding author’s e-mail: 2013101@hlju.edu.cn

Abstract. In scientific research, bioinformatics, Internet applications, e-commerce and many other application fields, the amount of data is growing at an extremely fast rate. To analyze and utilize these huge data resources, it is necessary to rely on effective data analysis technology. Big data application is a process of mining effective information from big data by using data analysis methods, providing auxiliary decisions for users and realizing the value of big data. In the process of data calculation and analysis, we will often find that the data from the same source will show multiple function images in the same coordinate system, which will make the same variable correspond to multiple values in data prediction analysis. We call it multivalued function here. This paper provides a method to identify multivalued functions. By using the maximum information coefficient (MIC) theory proposed by David n. Reshef, and using the data sampling method to calculate the function identifiability of the data set, then, the spectral clustering method is used for recognition, segment and mark the images of different functions. Finally, the regression function equations with different marks are obtained by Gauss-newton iteration method. The results can be used for data prediction and analysis to assist decision-makers to make reasonable judgments.

1. Instruction

As early as the 1880s, Karl Pearson proposed the Pearson correlation coefficient theory inspired by Francis Galton [1], however, this theory can only measure the linear correlation of two variables, X and Y, but not the nonlinear correlation [2]. Later, Spearman put forward the Spearman's rank correlation coefficient theory in 1904. It is a nonparametric index to measure the dependence of two variables and using the monotonic equation to evaluate the correlation between two statistical variables, it does not have the property of measuring the correlation of data widely too. The maximum information coefficient (MIC) theory proposed by David n. Reshef in 2011[3]. It can discover not only the linear function relation between variables but also the nonlinear function relation. Therefore, MIC measurement can be used not only to compare the strength of the same correlation vertically but also to compare the strength of different correlation horizontally. The MIC theory provides a strong theoretical basis for data analysis of multivalued function relations. Therefore, in this paper, the maximum information coefficient (MIC) value is used to evaluate the correlation between data, and to determine whether the data sets are computable and can be further analyzed.

Cluster analysis has always been a hot topic in data mining [4]. However, despite the rich development of clustering algorithms, no clustering algorithm has a good universality in the face of the diversity of data structures presented by complex data sets, most of the current clustering algorithms have obvious limitations of the application environment. In this paper, a spectral clustering algorithm based on spectral graph theory is applied. At present, many spectral clustering models and algorithms have appeared, such as a PF algorithm proposed by p.perona and w. t. feewman [5] and the
SM algorithm proposed by Shi and Malik [6]. Although there have been many spectral clustering algorithms, the only difference lies in the different matrices handled. Compared with the traditional clustering algorithm, it has the advantages of clustering in the sample space of arbitrary shape and convergence to the global optimal solution, and it can perform the better cutting of different functions in the same data set.

The selection of the regression model is also very important in data analysis. The commonly used regression models include linear regression, logistic regression, decision tree regression, support vector machine regression, etc. Gauss-newton iteration method [7] has a good fitting effect on data regression. It fits nonlinear data very well. However, all regression methods can not solve the problem of multivalued function analysis. We propose a novel multivalued function recognition algorithm based on spectral clustering. Experimental results show that our method has high precision.

2. Related Work

2.1. Definition of multivalued function
Multivalued function definition: Let’s say that X is a non-empty number set, and Y is a non-empty number set, F is a corresponding rule. If there is at least one element x in X, according to the corresponding rule F, Y has at least two elements that correspond to x. And for all the elements x in X, according to the corresponding rule F, there are elements y in Y that corresponds to X, then F is called a multivalued function from X to Y, write y = F (x).

According to the definition, we can further elaborate on F here. Here, we define F = \{f_1, f_2, ..., f_n\}, where \(f_1, f_2, ..., f_n\) represents each function relationship we segment and fit from the data set. The purpose of this paper is to solve \(f_1, f_2, ..., f_n\) on the data set.

2.2. Maximal Information Coefficient, MIC
Mutual Information [8] is a very useful Information measure in Information theory, which can be regarded as the Information contained in one random variable about another random variable, or the uncertainty reduced by knowing another random variable.

Definition of mutual information: Let's set two random variables (X; Y), their joint distribution is \(p(x, y)\), and the marginal distribution is \(p(x), p(y)\), mutual information I(X; Y) is the relative entropy of the joint distribution \(p(x, y)\) and the marginal distribution \(p(x)p(y)\). Namely:

\[
I_{x,y} = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right)
\]  

Calculation of maximum information coefficient: Record the set of variables x and y as D, D = \{(a_1, b_1), ..., (a_n b_n)\}, and it contains n points. The scatter plot is drawn with a as the x-coordinate and b as the y-coordinate, let \(G_{x,y}\) represent a grid divided by x × y. Let \(p_0, p_1, p_2, ..., p_x\) be the dividing point on the x-coordinate. Similarly, \(q_0, q_1, q_2, ..., q_y\) represent the segmentation points on the y-coordinate. Change the \(p_0, p_1, p_2, ..., p_{x-1}\) and \(q_0, q_1, q_2, ..., q_{y-1}\) value, can get x × y split approach. Let \(\text{Max}\{I_{x,y}\}\) represent the maximum mutual information value under different segmentation modes on \(G_{x,y}\), and finally normalize the maximum mutual information obtained to get the information normalized value \(M_{x,y}\):

\[
M_{x,y} = \frac{\text{max}\{I_{x,y}\}}{\log(\text{min}\{x,y\})}
\]

Where, \(M_{x,y}\) is the standardized value of the maximum information coefficient, \(I_{x,y}\) is the mutual information value in partition mode, \(\text{min}\{x,y\}\) is the minimum value between x and y. The coefficients of \(M_{x,y}\) are all within the range of (0,1), and the maximum value of \(M_{x,y}\) is MIC. Where x × y < \(n^{0.6}\), n is the number of data.
2.3. Spectral Clustering

Definition of Undirected weight graph: First define a graph $G = (V, E)$, where $V = \{v_1, v_2, ..., v_n\}$ represents the collection of points, $E = \{(v_i, v_j) | v_i, v_j \in V, v_i \neq v_j\}$ represents the graph of a collection of connected to the side, then we define the weight of the two edges connected in the graph $v_i, v_j$ as $w_{i,j}$. Since this is an undirected graph, $w_{i,j} = w_{j,i}$.

$$w_{i,j} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right), & i \neq j \\ 0, & i = j \end{cases} \tag{3}$$

Where: $\sigma$ is the scale parameter, $\sigma$ size in this paper was selected according to the thought of "local Scaling" proposed by Zelnik-Manor and Perona in the self-tuning spectrum clustering method [9].

For the two points $v_i$ and $v_j$ with edge connection, $w_{i,j} > 0$, for the two points $v_i$ and $v_j$ without edge connection, $w_{i,j} = 0$. For any point $v_i$ in the graph, its degree $d_i$ is defined as the sum of the weights of all the edges connected to it, namely:

$$d_i = \sum_{j=1}^{n} w_{i,j} \tag{4}$$

Definition of degree matrix: It is a matrix of $n \times n$, only the main diagonal has a value, corresponding to the degree of the $i$th point in the $i$th row, which is defined as follows:

$$D = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & d_n \end{bmatrix} \tag{5}$$

Definition of similarity matrix: For the input data set with $n$ data points, $V = \{x_1, x_2, ..., x_n\}$ is used to construct the similarity matrix $W \in R^{n \times n}$ according to the similarity between any two points.

$$W = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n} \\ w_{2,1} & w_{2,2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & w_{n-1,n} \\ w_{n,1} & \cdots & w_{n,n-1} & w_{n,n} \end{bmatrix} \tag{6}$$

Laplace matrix:

$$L = D - W \tag{7}$$

Where: $L \in R^{n \times n}$

For any set of two subgraph points $A, B \subseteq V, A \cap B = \emptyset$, we define the graph cutting weight between them:

$$W(A, B) = \sum_{i \in A, j \in B} w_{i,j} \tag{8}$$

For the set of $k$ subgraph points: $A_1, A_2, ..., A_k$, the RatioCut method [10] is adopted here and its definition is:

$$\text{RatioCut}(A_1, A_2, ..., A_k) = \frac{1}{2} \sum_{i=1}^{k} \frac{W(A_i, \overline{A_i})}{|A_i|} \tag{9}$$

Where: $|A_i|$ is the number of data points in set $A_i$, and $\overline{A_i}$ is the complementary set to $A_i$.

In this way, not only the minimization of RatioCut is considered for each cut, but also the maximization of the number of subgraphs is considered to avoid the extreme situation caused by the minimal cut.

Introducing an indicator vector, suppose $n$-dimensional vector $f_i \in \{f_1, f_2, ..., f_k\}$, $n$ is the number of data and $k$ is the number of clusters. Each element of the indicator vector represents the category attribution of the current data points. Each element of the indicator vector is defined as:
\[ f_{i,j} = \begin{cases} 
0, & v_j \notin A_i \\
\frac{1}{\sqrt{|A_i|}}, & v_j \in A_i 
\end{cases} \]  
(10)

where: \( j = 0, 1, \ldots, n \).

Vector \( f \) satisfies the following equation:

\[ f^T L f = |V| \text{RatioCut}(A_1, A_2, \ldots, A_k) \]  
(11)

Then the \( \text{RatioCut} \) function can be optimized as:

\[ \text{RatioCut}(A_1, A_2, \ldots, A_k) = \sum_{i=1}^{k} f_i^T L f_i = \text{tr}(H^T L H) \]  
(12)

Where: \( |V| \) denotes the number of data sample points, \( H \) is a matrix composed of indicator vectors \( f_i \). \( \text{tr}(H^T L H) \) is the trace of matrix.

Therefore, the cut graph optimization goal of spectral clustering is transformed into:

\[
\begin{cases}
\min \text{tr}(H^T L H) \\
\text{s.t. } H^T H = I
\end{cases}
\]  
(13)

This is a standard solution to trace minimization problem. According to the properties of the Laplace matrix, the optimal solution of this optimization problem can be converted into finding the eigenvectors corresponding to the \( k \)-1 minimum eigenvalues in the front of \( L \) matrix, and each eigenvector indicates a different cluster data set.

2.4. Gauss-Newton iteration method

Gauss-Newton algorithm to solve nonlinear equations: Let \( h \) be the direction of \( F(x) \) descent, and \( h^TF'(x) < 0 \), for any small enough \( \alpha > 0 \), \( F(x + \alpha h) < F(x) \). Let \( f_i(x) \) be a non-linear function, so that the vector-valued function is as follows:

\[ F(x) = \frac{f(x)^T f(x)}{2} \]  
(14)

than:

\[ F(x_k + h) \approx \frac{f(x_k + h)^T f(x_k + h)}{2} = F(x_k) + h^T J_k f_k + \frac{h^T J_k f_k h}{2} \]  
(15)

where: \( J_k \) is the Jacobian matrix of the \( k \)-th iteration, there are \( N \times M \) elements, and the expression of each element is:

\[ J_{ij} = \frac{\partial f_n(x_k)}{\partial x_m} \text{ (} i = 1,2,\ldots,N; j = 1,2,\ldots,M \text{)} \]  
(16)

In order to minimize \( F(x_k + h) \), let \( F(x_k + h) \) be the first derivative of \( \frac{\partial F(x_k h)}{\partial h} = J_k f_k + J_k h \) so that \( \frac{\partial F(x_k h)}{\partial h} = 0 \). According to the nonlinear least squares objective function expression, we can get:

\[ \nabla h = -(J_k f_k)^{-1} J_k f(x^k) \]  
(17)

\[ x_{k+1} = x_k + \nabla h \]  
(18)

\[ \nabla F(x) = F(x_{k+1}) - F(x_k) \]  
(19)

2.5. MKSR algorithm for multivalued function recognition

Considering that the data set is generally large, to reduce the time complexity and achieve faster calculation results. The MIC value was calculated from random sampling data, \( \text{MIC} \geq 0.4 \) (0.4 is an empirical value) [11] is used to determine whether the data set is continuously computable. The time complexity of spectral clustering is positively correlated with the data quantity. First, we used the k-means method to carry out rough clustering for the data, the data center points of rough clustering are used to classify spectral clustering. To sum up, this greatly reduces the time complexity of the algorithm.

MKSR algorithm as follows:
Algorithm: MKSR

Input: $X = \{x_1, x_2, ..., x_N\}$
Output: $F = \{f_1, f_2, ..., f_j\}$

1. Take $m$ samples from input data $X$ by random sampling, calculate MIC
2. if $\text{MIC} \geq 0.4$ do:
3. aggregate samples into $i$ categories by K-Means, mark center as $\{c_1, c_2, ..., c_i\}$
4. calculate $W_{ij}$ from Eq. (3)
5. calculate $D$ by equation (5), $L = D - W$
6. calculate $\xi = \{\xi_1, ..., \xi_{k-1}\}$ s.t. $L\xi = \lambda_1\xi_1, \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_i$
7. calculate classification mark $\{G_1, G_2, ..., G_j\}$ by $\xi$
8. calculate $F = \{f_1, f_2, ..., f_j\}$ by Gauss-newton algorithm
9. return $F$

3. Experiment

3.1. Simulation data results

Figure 1. DBSCAN algorithm

Figure 2. MKSR algorithm

Figure 3 and Figure 4 are the comparison of DBSCAN algorithm and MKSR algorithm.

The above comparison shows that the DBSCAN algorithm cannot cut dense clusters with multiple branches, and its function fitting results have no reference significance. By contrast, obviously, the MKSR algorithm, which has a good effect on graph cutting by spectral clustering, can accurately identify and regression multivalued functions.

3.2. Real data set results

This data is the data visualization result calculated by MKSR algorithm for the closing price (X) and trading volume (Y) of the Shanghai Stock Exchange in the third and fourth quarters of 2018[13].

Figure 3. The closing price is between 0 and 60 with DBSCAN algorithm

Figure 4. The closing price is between 0 and 60 with MKSR algorithm

By comparing the two algorithms with the real stock data, it is proved that the MKSR algorithm we proposed has incomparable advantages in multivalued function recognition. Of course, the MKSR algorithm can run on many data sets. This paper takes stock data as an example to prove this algorithm.
4. Conclusion
In this paper, the advantages of MIC in detection data correlation and spectral clustering in graph cutting are utilized to propose the MKSR algorithm combined with the respective advantages of the k-means method and Gauss-newton iterative algorithm. Firstly, MIC values obtained from data sampling are used to predict the computability of data, then, the k-means method is used for initial clustering of large data sets to obtain the data center points of each class, and the data center points are used for spectral clustering division marking. Finally, the Gauss-Newton iterative algorithm is used to solve the nonlinear function relations for different labels, so as to realize the recognition and output of multivalued functions.

References
[1] Benesty J, Chen J, Huang Y, et al. Pearson Correlation Coefficient[M]// Noise Reduction in Speech Processing. 2009.
[2] Wang J. Pearson Correlation Coefficient[M]. 2013.
[3] Reshef D N, Reshef Y A, Finucane H K, et al. (2011) Detecting Novel Associations in Large Data Sets[J]. Science, 334 (6062):1518-1524.
[4] Borcard D, François Gillet, Legendre P. (2018) Cluster Analysis[M]// Numerical Ecology with R.
[5] Perona P. (1998) A factorization approach to grouping[J]. Proc. ECCV.
[6] Shi J, Malik J. (2000) Normalized Cuts and Image Segmentation[J]. IEEE Trans.pattern Anal.mach.intell, 22(8):888-905.
[7] Giftthaler M, Neunert M, Stäuble M, et al. (2017) A Family of Iterative Gauss-Newton Shooting Methods for Nonlinear Optimal Control[J].
[8] [Belghazi M I, Baratin A, Rajeswar S, et al. (2018) MINE: Mutual Information Neural Estimation[J].
[9] Zelnik-Manor L. (2004) Self-tuning spectral clustering [J]. Advancesin Neural Information Processing Systems,17: 1 601-1608.
[10] Hagen L, Kahng A B. (1992) New spectral methods for ratio cut partitioning and clustering[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits, and Systems, 11(9):1074-1085.
[11] Iuchi H, Sugimoto M, Tomita M. (2018) MICOP: Maximal information coefficient-based oscillation prediction to detect biological rhythms in proteomics data[J]. Bmc Bioinformatics, 19(1):249.
[12] Shanghai Stock Exchange .(2018).http://www.sse.com.cn/