INFLUENCE POWER-BASED CLUSTERING ALGORITHM FOR MEASURE PROPERTIES IN DATA WAREHOUSE

Min Ji [a, *], Fengxiang Jin [a], Ting Li [a], Xiangwei Zhao [a], Bo Ai [a]

[a] Geomatics College, Shandong University of Science and Technology, 579 Qianwangang Road, Economic & Technical Development Zone, Qingdao, China, 266510 – jimin@sdust.edu.cn

KEY WORDS: Influence Power, Hierarchical Tree, Neighbor Function Clustering, Data Mining, Gravitational Clustering, Nature Clustering

ABSTRACT:

The data warehouse's fact table can be considered as a multi-dimensional vector point dataset. In this dataset, each point's measure property can be transformed as the influence power against its neighbor points. If one point's measure is larger, it would have more influence power to attract its neighbor points, and its neighbors would have a trend to be absorbed by this point. Being inspired by the Gravitational Clustering Approach (GCA), the paper introduces a new method named IPCA (Influence Power-based Clustering Algorithm) for clustering these vector points. The paper first defines several concepts and names the local strongest power points as Self-Strong Points (SSPs). Using these SSPs as the initial clustering centers, IPCA constructs serials of hierarchical trees which are rooted by these SSPs. Because there are only a few SSPs left, by using each SSPs' influence power, the paper adopts the neighbor function clustering method to define the clustering criteria function, and gives the detail clustering procedure of SSPs. IPCA follows the nature clustering procedure at the micro-level, with a single scan, it can achieve the initial clustering. From the experiment result, we can see that IPCA not only identifies different scale clusters efficiently, but it also can get arbitrary shape clusters easily.

1. INTRODUCTION

Clustering is of fundamental importance in machine learning and data mining (Kundu, 1999). The principles of clustering include partitioning, hierarchical based, grid based, and model based (Guha, 2000. Dutta, 2005). Most of these clustering algorithms use the distance to measure the similarity between two vector points. Sometime this can partition a big nature cluster into several sub-clusters. The Gravitational Clustering Approach (GCA) (Giraud, 2005. Mohayaee, 2003. Chen, 2005) quotes the universal gravitation principle, and calculates the attraction force between two points to measure whether they merge or not. GCA doesn't restrict the clustering radius (Jiang, 2005). According the gravity power, GCA can partition vector points into a few super-spheres which have different radius, and can hold the nature clustering procedure.

As you know, the power of the gravity between two objects is determined by their qualities. If one object has more quality, it would have more power to attract its neighbor objects, and the neighbors would have a trend to be absorbed by this object. For the measure property in data warehouse's fact table, we can consider it as the purpose or result of data observation or statistics. It will have more roles and be more important than other dimensional attributes in the clustering procedure. The fact table can be considered as a multi-dimensional vector dataset, and one record is a vector point. According to the GCA idea, the magnitude of one point's measure property will determine its influence power against its neighbor points or its dependent direction. Based on this, we propose a new clustering algorithm –Influence Power-based Clustering Algorithm (IPCA). IPCA can achieve the initial clustering with a single scan. It starts from the micro-level, and can satisfy the nature clustering procedure. The rest of the paper is organized as follows; Section 2 defines several core concepts and discusses their usage in IPCA. Section 3 describes the algorithm's whole clustering idea, and defines the SSPs' clustering procedure which is referenced the neighbor function clustering method. Section 4 gives the calculation steps and Section 5 describes a simple by using a multi-dimensional fishery dataset. The result shows that IPCA is efficient, and can identify different-scale clusters or arbitrary shape clusters.

2. BASIC CONCEPTS OF IPCA

With the long-term observation and statistics of the natural and social phenomena, the people has accumulated a large number of multi-dimensional datasets, such as biological population distribution data, natural resource survey data, economic development statistics data, etc. All these data are integrated respectively in data warehouse's multi-dimensional fact tables (Marc, 1997. Coliat, 1996. Han, 2000). In the fact table, there are two kinds of attributes: dimensions and measures. We can classify these dimensions as classification dimension, order dimension, temporal dimension, spatial dimension, and so on. All these dimensions construct a multi-dimension space, and divide the space as a multi-dimension cube collection. Each record in the fact table corresponds one of these cubes. While measure properties correspond to the observation value or the statistics in the cube, and represent the purpose and results of data acquisition. They should have more roles in the clustering procedure. In order to descript more clearly for the follow sections, we gives these follow concepts.

The Influence Power is used to measure the important degree of a multi-dimension point. For the multi-dimension dataset in data warehouse, it refers to the measure property of one point. According to GAC idea, if one point has a higher measure value,

* Min Ji , Associate Professor of Shandong University of Science and Technology, Research Fields: Spatial Data Organization and GIS Application Developing, Email: jimin@sdust.edu.cn, jamesjimin@hotmail.com

it will have a higher influence power to attract its neighbor points, and it will have more possibilities to become a cluster center. So, we use the magnitude of one point's measure property as the influence power, and also normalize it by using Eq. (1).

$$I_i = \frac{p_i - P_{\min}}{P_{\max} - P_{\min}} \qquad (i = 1,2,\cdots,n) \qquad (1)$$

where n = the number of all points
p_i = the measure value of the ith point
$P_{\max}$ = the maximum measure value
$P_{\min}$ = the minimum measure value
$I_i \in [0,1]$

Grid Neighbor Points (GNPs) refer to those points that have one unit distance with the current point along each dimension in the multi-dimensional cube collection (also called as grid matrix). Fig. 1 shows a three-dimensional cube, the red point at the center is the current point that we will examine its influence power, and all the black points are its GNPs.

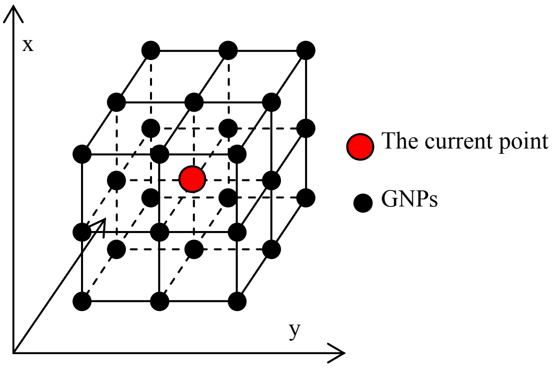

Figure 1. The relationship of current point and its GNPs

The Maximum Connection Strength point (MCS point) is the point among all GNPs of the current point which has the strongest influence power. If there are several GNPs which influence power are larger than the current point's, the current point will have a trend to be absorbed by these GNPs. Which GNP should the current point belong to? We should follow the principle of playing up to those in power. And then the current point should connect to the GNP which has the strongest influence power. That will be the maximum connection strength point.

The Strongest Power Connection direction (SPC direction) is the current point's attaching direction that connects to its MCS point.

Self-Strong Point (SSP) is the point that has the maximum influence power at local area. If one point's influence power is higher than any one's power of its GNPs, it will become the local strongest power point. Because it has no MCS point, we call it as self-strong point. There is a special situation. If one point has no GNPs, in another words, it is an isolate point, we also call it as SSP. SSPs can be used as the initial clustering centers.

SuBsidiary Points (SBPs) refer to those points that are not SSPs. Each SBP has its own MCS point, and will connect to its MCS point. But one SBP's status is not everlasting, it also can be a MCS point, and can have its own SBPs.

Natural Clustering Hierarchy tree (NCH tree) is the core concept. NCH tree is composed by a SSP, MCS points and SBPs. The root node is the SSP. Each parent node corresponds to a MCS point, and the SSP is the top MCS point. Each MCS point at least has one child node, and this child is its SBP. One SBP also can be a MCS point, and can have its own SBPs, so there forms one to many hierarchical relationships between MCSs and SBPs. One NCH tree can be seen as an initial cluster.

3. THE IPCA CLUSTERING ALGORITHM

3.1 The Initial Clustering

According to the description of Section 2, the measure property can be as a key index for estimating one multi-dimension point's important degree. If one point's measure value is higher, it will have more influence power to attract its GNPs, and it will have more possibilities to become a cluster center. Based on this common knowledge, by calculating each point's influence power, we can get many local strongest power points (SSPs), and can select these SSPs as the initial clustering centers. SSPs can be as the first level of MCS points, and the next step is just continuously expanding MCS points' SBPs. Because one SBP can also be a MCS point, only after each point has been scanned, and has found its MCS point, the initial clustering would accomplish. After this procedure, we will get a few NCH trees. The number of NCH trees just equals the SSPs'. If the condition is good enough, we would get the final clustering result. But for a general condition, we should go another deeper step.

After the initial clustering procedure, each NCH tree represents a sub-cluster. Our follow issue is just how to measure the similarity between these sub-clusters. Because each NCH tree has a SSP, we only measure the similarity between SSPs, we will get the answer. Section 3.2 just describes SSPs' clustering procedure.

3.2 SSPs' Clustering Algorithm

In accordance with the whole clustering idea, the final clustering process is mainly reflected in SSPs' clustering procedure. The quality of SSPs' clustering will have a direct impact on the quality of the final clustering result. Inspired from the neighbor function clustering method (Sun, 2001), we proposed the follow influence power-based SSPs' neighbor function clustering method.

Firstly, we give a distance threshold $D_{\max}$ which is calculated by numerical dimensions of these SSPs. Within one SSP's $D_{\max}$ rang, there forms a SSPs' sub-group. Each SSP in the sub-group will have an attraction force on this current point. We can use Eq. (2) to calculate the force.

$$E_{oj} = I_j / D_{oj} \qquad (j = 1,2,\cdots,m) \qquad (2)$$

where m = other SSPs within current SSP's D_{max} rang
I_j = the influence power of the jth SSP
D_{oj} = the distance between current SSP and its jth SSP

According to E_{oj} values, we sort them descending and there forms the influence neighbor order between current SSP and its surrounding SSPs. Here, we give another concept – the Influence Neighbor Points (INPs), which are those SSPs within current SSP's D_{max} rang.

For any two SSPs P_i and P_j, if P_i is P_j's Ith INP, we designate I as the influence neighbor coefficient that P_i puts on P_j, denoted as E (i, j) = I; Similarly, we can get E (j, i) = J. Then, we define Eq. (3) as the influence neighbor function between P_i and P_j.

$$\alpha_{ij} = E(i, j) + E(j, i) - 2 = I + J - 2 \tag{3}$$

If P_i and P_j are the first INP for each other, then $\alpha_{ij} = 0$. This shows that the smaller the α_{ij}, the greater the attraction between two points, and the more possibility to cluster together. Suppose the number of SSPs is N, then $\alpha_{ij} \le 2N - 4$. If the distance between P_i and P_j exceeds D_{max}, we require $\alpha_{ij} = 2N$. In order to avoid SSP's self-loop clustering, we also require $\alpha_{ii} = 2N$.

In SSPs' clustering process, if P_i and P_j can merge together, we can claim that P_i and P_j are connected to each other. In order to show the loss for the connection, we introduce the SSP Connection Loss concept. According to Eq. (3), we can use α_{ij} as the SSP Connection Loss.

If SSPs can merge together, we define the connection loss in this cluster as $\Sigma\alpha_{ij}$. If there are c clusters: ω_p, p = 1,2, ..., c, we can define the total connection loss among these clusters as Eq. (4).

$$L_w = \sum_{p=1}^{c} \sum_{\substack{P_i \in \omega_p \\ P_j \in \omega_p}} \alpha_{ij} \tag{4}$$

In order to describe SSPs' internal expanding degree, we define the inner maximum connection loss in one cluster as Eq. (5).

$$\alpha_{pm} = \underset{\substack{P_i \in \omega_p \\ P_j \in \omega_p}}{MAX} [\alpha_{ij}] \qquad p = 1,2,..., c \tag{5}$$

In order to describe the similarity between clusters, we define the connection loss between clusters as Eq. (6).

$$\gamma_{pq} = \underset{\substack{P_i \in \omega_p \\ P_j \in \omega_q}}{MIN} [\alpha_{ij}] \qquad p, q = 1,2,..., c; p \ne q \tag{6}$$

In order to determine the quality of the previous iterative clustering results, we define the minimum connection loss between cluster ω_p and all the other clusters as Eq. (7).

$$\gamma_{pk} = \underset{\substack{q \\ q \ne p}}{MIN} [\gamma_{pq}] \quad q = 1,2,..., c \tag{7}$$

If $\gamma_{pk} > \alpha_{pm}$ $\quad and \quad$ $\gamma_{pk} > \alpha_{km}$, it shows that the previous clustering result is successful, otherwise the clusters should merge together. The merge condition is as $\gamma_{pk} \le \alpha_{pm} \quad or \quad \gamma_{pk} \le \alpha_{km}$.

In order to describe the total loss between all the clusters, we define Eq. (8) as the loss-cost function between clusters.

$$\beta_p = \begin{cases} (\alpha_{pm} - \gamma_{pk}) + (\alpha_{km} - \gamma_{pk}) & if\, \gamma_{pk} > \alpha_{pm}\, and \gamma_{pk} > \alpha_{km} \\ \alpha_{pm} + \gamma_{pk} & if\, \gamma_{pk} \le \alpha_{pm}\, and \gamma_{pk} > \alpha_{km} \\ \alpha_{km} + \gamma_{pk} & if\, \gamma_{pk} > \alpha_{pm}\, and \gamma_{pk} \le \alpha_{km} \\ \alpha_{pm} + \alpha_{km} + \gamma_{pk} & if\, \gamma_{pk} \le \alpha_{pm}\, and \gamma_{pk} \le \alpha_{km} \end{cases} \tag{8}$$

In Eq. (8), the first case is a reasonable clustering result; β_p is negative, indicating that the loss is negative. The other cases are unreasonable, there has a need to merge clusters, β_p is positive, indicating there has some connection loss.

Based on Eq. (8), we define the total connection loss among all the clusters as Eq. (9).

$$L_B = \sum_{p=1}^{c} \beta_p \tag{9}$$

The final goal of SSPs' Clustering is to enable γ_{pk} as large as possible, and enable α_{pm} as small as possible, so we define the clustering criterion function for SSPs as Eq. (10).

$$J_L = L_W + L_B \quad \rightarrow \quad Min \tag{10}$$

4. IPCA CLUSTERING PROCEDURE

Based on the description and definition of the earlier sections, we experimented with a three-dimension dataset whose record number is 2187. We present serials of efficient steps for obtaining the finial clusters by using each point's measure property. All the steps are as follows.

Step 1. Find the maximum measure value P_{max} and the minimum value P_{min} from the entire dataset.

Step 2. Normalize the measure value for each record by using P_{max} and P_{min}, and get each point's influence power I_i.

Step 3. According to each point's influence power, select SSPs and construct NCH trees. From the 2178 points, we got 153 SSPs.

Step 4. Sort these SSPs descending by using their influence power, and construct the SSPs' dataset {P_1, P_2, …, P_N}.

Step 5. Calculate the Euclidean inverse distance matrix ***D***, the element in ***D*** is as D_{ij} = 1/ d (P_i, P_j), where d (P_i, P_j) is the Euclidean distance between the *i*th SSP and the *j*th SSP, i ≠ j. Set the main diagonal element as 0.

Step 6. According to the data characters, set the distance threshold D_{max}. Set the matrix element's value as 0 if its value is smaller than $1/D_{max}$.

Step 7. Construct the influence power vector $\boldsymbol{I} = (I_1, I_2, ..., I_N)$.

Step 8. Calculate the attraction degree matrix $\boldsymbol{M}=\boldsymbol{DI}^T$.

Step 9. Calculate the influence neighbor coefficient of each non-zero element in matrix ***M***, and form the influence neighbor coefficient matrix ***H***.

Step 10. According to the non-zero element in matrix ***H***, calculate the influence neighbor function matrix ***L***, where the element $L_{ij} = h_{ij} + h_{ji}$-2. Set all zero elements as 2*N*. L_{ij} represent the connection loss if two SSPs merge together.

Step 11. Select these first-M SSPs as the clustering centers. According matrix ***L*** to determine whether there are any SSPs which have minimum influence neighbor function value (min-value) for each other in these first-M SSPs, if existing, then merge these SSPs together. After that, determine whether the remaining SSPs have min-value with these first-M SSPs, if existing, then merge them to their corresponding clusters, otherwise, they will be consider as a single cluster. These single clusters maybe outliers, if one single cluster doesn't have SBPs, we should delete it from the cluster collection.

Step 12. According Eq. (5-7) to calculate $\gamma_{pk}(p=1,2,\cdots,c)$、$\alpha_{pm}$、$\alpha_{km}$, if $\gamma_{pk} \le \alpha_{pm}$ *or* $\gamma_{pk} \le \alpha_{km}$, merge ω_p and ω_k into one cluster, and then repeat this step. This procedure will iterative until Eq. (10)'s J_L value is not diminished

Step 13. After accomplish SSPs clustering procedure, finally add each node of each NCH trees to its corresponding cluster and finish the whole clustering procedure.

5. EXPERIMENTS

Based on the previous clustering procedure, by using 2178 records in one marine fishery data warehouse's fact table, we got 12 natural clusters, which include 7 big clusters, 2 small clusters and 3 outlier clusters, just as shown in Table 1. From Table 1, we can see that cluster 10, 11, 12 only have one SSP , and do not have any SBPs, they are far from any clusters, obviously they are outliers, we should delete them from the finial clustering result. Cluster 6 and cluster 9 only have a small number of SSPs and their total point numbers are also far less than other clusters. They can be considered as small-size clusters. Cluster 2 is the biggest-size cluster. It has 40 SSPs, and 722 points, more than 30 times of cluster 6 or cluster 9. This demonstrated that IPCA method can identify multiple clusters with different scales very clearly. Fig. 2 illustrates the clustering distribution in three-dimensions which are x-dimension, y-dimension, and T-dimension. The x, y values represent the vector point's coordinate position in our flat space, and the T-value represents the point's happening time in the temporal space. From the three-dimensional clustering distribution map, we can see that there are very large changes in these clusters' shapes. They are not spherical, not liner. They can be arbitrary shapes. This will be an excited character than the traditional clustering algorithm.

No.	SSPs Num.	Total Num.	Measure Property	Cluster Center		
				x	y	T
1	25	482	1661.6	146	41	47
2	40	722	1619	154.5	43.5	35
3	15	239	1002	166.5	41.5	29
4	23	317	771.5	159.5	44	38
5	18	145	751.3	161.5	41.5	31
6	2	24	600.6	151.5	41	47
7	11	130	437.9	-175	41	27
8	10	95	219.8	178	41	28
9	6	21	34.7	171.5	38.5	24
10	1	1	8.7	166	38	20
11	1	1	8.1	178.5	40.5	39
12	1	1	7	180	36.5	26

Table 1 The clustering result table of 2178 vector points

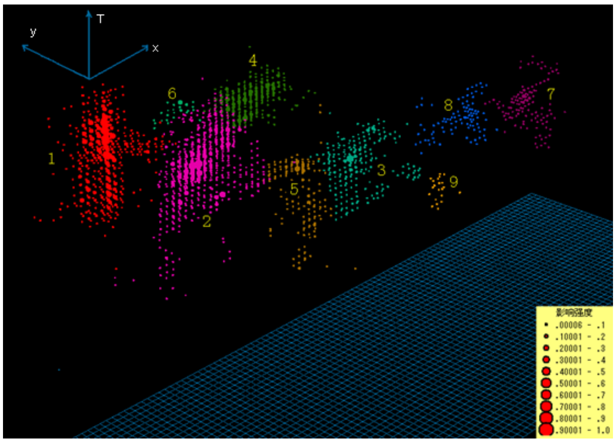

Figure 2 The 3D distribution map of clustering result

6. CONCLUSION

This paper presents the IPCA algorithm, a new clustering algorithm based on objects' influence power against each other, which is inspired from the gravity theory in physics. This method is particularly suitable for handling multi-dimensional huge data collections in data warehouse's fact table. Because measure properties are observation values or statistic results, they should have more roles in data mining procedure. IPCA just uses the measure property of each multi-dimensional record to measure its attraction force on its neighbor records in the multi-dimensional space. Only for a single scan, this algorithm can get the initial clustering result and construct many hierarchy trees which are rooted by self-strong points which have the strongest influence power in the local area. If the condition is good enough, we can get the finial clusters. IPCA follows the natural clustering process, and the experimental results also show that it can identify any size clusters and arbitrary shape clusters efficiently. These two characters will make it as a new member in the data clustering analyze family.

REFERENCES

S. Kundu, 1999. Gravitational Clustering: a new approach based on the spatial distribution of the points, Pattern Recognition, 32, pp. 1149-1160.

S. Guha, R. Rastogi, and K. Shim, 2000. Rock: A robust clustering algorithm for categorical attributes, Information systems, 25(5), pp. 345-366.

M. Dutta, A. Kakoti Mahanta, A. K. Pujari, 2005. QROCK: A quick version of the ROCK algorithm for clustering of categorical data, Pattern Recognition, 26, pp. 2364-2373.

J.A. Garcia, J. Fdez-Valdivia, F.J. Cortijo, and R. Molina, 1995. A dynamic approach for clustering data, Signal Processing, 44,pp. 181-196,

C. Giraud, 2005. Gravitational clustering and additive coalescence, Science Direct, 115, pp. 1302-1322.

R. Mohayaee, L. Pietronero, 2003. A cellular automaton model of gravitational clustering, Science Direct, 323, pp. 445-452.

C.Y. Chen, S.C. Hwang, Y.J. Oyang, 2005. A statistics-based approach to control the quality of subclusters in incremental gravitational clustering, Pattern Recognition, 38, pp. 2256-2269.

S.Y. Jiang, Q.H. Li, 2005. Gravity-based clustering approach, Journal of Computer Applications, 2, pp. 285-300.

Jiawei Han, M. Kamber, 2000. DATA MINING: Concepts and Techniques, Morgan Kaufmann Publishers, pp. 100-150.

Marc G., 1997. A Foundation for Multi-dimensional Databases, In Proc. of the 23rd VLDB Conference, pp. 106-115.

Coliat G, 1996. OLAP, relational, and multi-dimensional database system, ACM SIGMOD Record, 25(3), pp. 64-69.

J.X. Sun, 2001. Modern Pattern Recognition, National University of Defense Technology, pp. 40-43.

APPENDIX

This article was sponsored by Chinese National Programs for High Technology Research and Development (2009AA12Z147); the Key Laboratory of Surveying and Mapping Technology on Island and Reed, State Bureau of Surveying and Mapping (2009B14); the Key Laboratory of Advanced Engineering Surveying of SBSM (TJES0805).