Online Rule Updating System Using Evolutionary Computation for Managing Distributed Database

Wirarama Wedashwara, Shingo Mabu, Masanao Obayashi and Takashi Kuremoto
Graduate School of Science and Engineering, Yamaguchi University
Tokiwadai 2-16-1, Ube, Yamaguchi 755-8611, Japan
E-mail:{ t001we, mabu,m.obayas, wu}@yamaguchi-u.ac.jp

Abstract

This research proposes a decision support system of database cluster optimization using genetic network programming (GNP) with on-line rule based clustering. GNP optimizes cluster quality by reanalyzing weak points of each cluster and maintaining rules stored in each cluster. The maintenance of rules includes : 1) adding new relevant rules, 2) moving rules between clusters and 3) removing irrelevant rules. The simulation focuses on optimizing cluster quality response against several unbalanced data growth to the data-set that is working with storage rules. The simulation results of the proposed method shows better results compared to GNP rule based clustering without on-line optimization.

Keywords: Genetic Network Programming, Rule Based Clustering, Cluster Optimization

1. Introduction

Nowadays many large scale database systems with very high data growth are being utilized to improve the global human activity, such as communication, social networking, transaction, banking, etc. A distributed database management system becomes a solution to improve data access speed by organizing data in multiple storages for multiple types of user accesses. Problems of distributed database management system are not only how to manage the number of data, but also how to organize data patterns in distributed storages. Clever data organization is one of the best ways to improve the retrieval speed and reduce the number of disk I/Os and thereby reduce the query response time.

In this paper, we propose a decision support system for database cluster optimization using Genetic Network Programming (GNP) with on-line rule based clustering. Main purpose of this research is to provide an on-line algorithm to maintain the cluster adaptability against several unbalanced data growth. For example, the unbalanced data growth occurs when different kinds of items (data) comparing to the items stored in the current database begin to be stored as the time goes on (the trend of data is changed).

2. Review of the Proposed Framework

2.1. *Rule Based Clustering*

Rule based clustering is one of the solutions to provide automatic database clustering and interpretation of data storage patterns. Rule based clustering represents data patterns as rules by analyzing database structures on both of attributes and records[3,4].

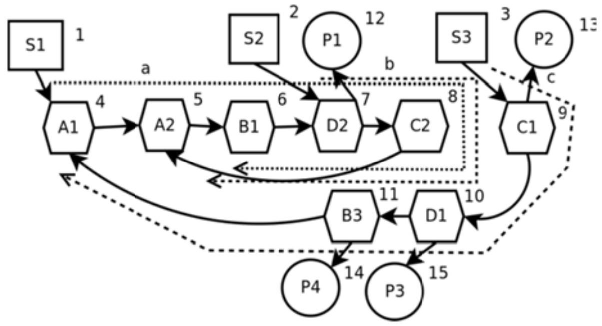

Fig. 1. GNP Implementation on Cluster Optimization.

Table 1. Gene Structure of GNP Corresponding to the Program in Fig. 2.

i	NTi	Ai	Ri	Ci
1	1	0	0	4
2	1	0	0	7
3	1	0	0	9
4	2	A	1	5
5	2	A	2	6
6	2	B	1	7
7	2	D	2	8,12
8	2	C	2	5
9	2	C	1	10,13
10	2	D	1	11,15
11	2	B	3	4,14
12	3	1	1	0
13	3	2	3	0
14	3	3	2	0
15	3	4	4	0

2.2. *Genetic Network Programming*

Genetic network programming (GNP) that is an evolutionary optimization technique with directed graph structures is used to provide data classification. GNP has a distinguished representation ability with compact programs, which is derived from the re-usability of nodes that is inherently equipped function of the graph structures. For the purpose of rule based clustering, GNP is useful to handle rule extraction from data-sets by analyzing the records.

2.3. *Silhouette*

Silhouette value is used to evaluate the clustering results. Silhouette provides a succinct graphical representation of how well each object lies within its cluster[5,8]. Silhouette value is calculated by Eq. 1.

$$
\begin{aligned}
s &= \frac{b-a}{\max\{a,b\}} \\
&= \begin{Bmatrix} 1-a/b & if\ a<b \\ 0 & if\ a=b \\ b/a-1 & if\ a>b \end{Bmatrix}
\end{aligned}
$$

s: Silhouette value for a single sample. The Silhouette value for a set of samples is given as the mean of the Silhouette values of each sample, a: mean distance between a sample and all the other points in the same cluster, b: mean distance between a sample and all the other points in the second nearest cluster. A good clustering result will make the silhouette value be close to 1.0 and a bad clustering result will close to -1.0. Silhouette value is used as a threshold in the rule updating process of GNP.

3. Detail of the Proposed Method

GNP optimizes the cluster quality by re-analyzing each cluster and updating rules stored in each cluster. The proposed method provides a rule updating mechanism as an additional function to maintain rules as shown in Fig 1. The rule updating mechanism includes: 1) adding new relevant rules; 2) moving rules between clusters and 3) removing irrelevant rules. The rule extraction and rule updating task is executed by the evolutionary optimization technique of GNP to get the best cluster adaptability against several unbalanced data growth.

GNP is used to extract rules from a data-set by analyzing all the records. Phenotype and genotype structures of GNP are described in Fig. 1 and Table 1, respectively. In Fig. 1, each node has its own node number (1–15), and in Table 1, the node information of each node number is described. The program size depends on the number of nodes, which affects the amount of rules created by the program. There are 3 kinds of nodes used in GNP rule extraction for the cluster rule updating. 1) Start node (rectangle) represents the start point of the sequence of judgment nodes which are executed sequentially by their connections. Multiple placements of start nodes will allow one individual to extract a variety of rules, which is shown in Table 1 2) Judgment node (hexagon) represents an attribute of the data-set, which is represented by Ai showing an index of attribute i such as price, stock, etc., and Ri showing a range index of attribute i. For example, Ai=A represents price attribute, and Ri=1 represents value range [0,50] and Ri=2 represents value range [51,80]. 3) Processing node (round) shows the end point of the sequence of judgment nodes and processes the rule updating in a cluster whose cluster number is described in the processing node . For example P1 processes an addition of extracted rules to cluster no 3. Multiple placement of processing nodes will make one individual extract variety of rules, which is shown in Table 2. Sequences of nodes starting from each start node (S_1,S_2,S_3) are represented by dotted line a, b and c. A node sequence flows until support for the next combination does not satisfy the threshold. The nodes with the attributes that have already appeared in the sequence will be skipped. Candidate rules extracted by the program of Fig. 1 to the data-set of Table 1 are shown in Table 2. In Table 2, three rules are extracted by the node sequence from each start node.

Fig. 2 shows and example of the Silhouette values of rules with only one attribute. In Fig. 3 threshold is set at 8.5, so only the attributes with Silhouette values under 8.5 will be used for the judgment nodes in GNP for rule extraction. In Fig. 2, attribute of price is not included in the rule extraction of GNP because its silhouette values are never under 8.5 (rule updating is not necessary), but the values of stock and weight for some data are under 8.5, thus those attributes are included in GNP for updating rules. Each cluster has dominant values of attributes which are anchors of cluster quality, which then influence the silhouette values. For example, when 10kg is a dominant value for the attribute of weight in a cluster, farther values from 10kg will have a lower silhouette and should be moved to another cluster to improve the cluster quality.

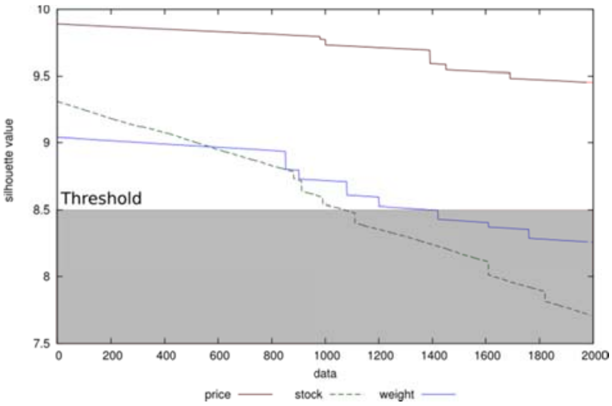

Fig. 2. Example of Silhouette Values

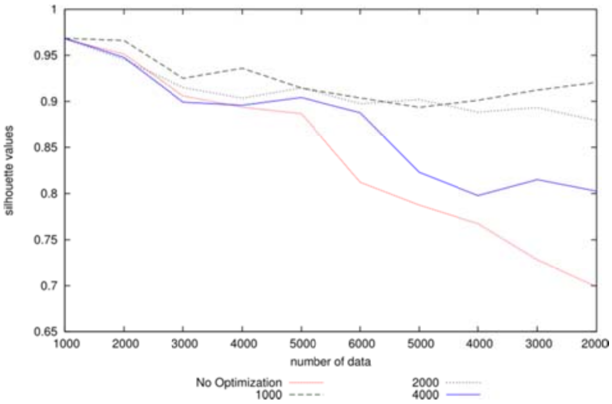

Fig. 3. Graph of Simulation Results of Rule Updating Frequency and Number of Data Comparison.

4. Simulation

The simulation focuses on verifying the cluster adaptability against several unbalanced data growth of the data-set, where cluster adaptability is evaluated by silhouette values. Data-set used in the simulations has 1000 data with 8 attributes. The adaptability is

Table 3. Simulation Result of Rule updating Frequency and Number of Data Comparison.

Step	Number of Data	Increment/Decrements	Rule updating frequency			
			No rule updating	1000	2000	4000
1	1000 (Default)	-	0.967	0.967	0.967	0.967
2	2000	+1000	0.945	0.965*	0.944	0.947
3	3000	+1000	0.902	0.923*	0.915*	0.899
4	4000	+1000	0.892	0.935*	0.902	0.895
5	5000	+1000	0.882	0.912*	0.909*	0.903*
6	6000	+1000	0.812	0.902*	0.897	0.887
7	5000	-1000	0.787	0.892*	0.901*	0.821
8	4000	-1000	0.765	0.901*	0.888	0.797
9	3000	-1000	0.723	0.912*	0.892*	0.815*
10	2000	-1000	0.698	0.909*	0.879	0.802
Average			0.832	0.938	0.923	0.8845

evaluated by the following two points. 1) The number of data in the database is changing as the time goes on. The initial number of data is 1000, then every time step, 1000 new data are added to the database. After the number of data reaches 6000, 1000 data are decreased every time step. 2) The rule updating frequency, where the rule updating is executed when a predefined number of new data are given to the data-set (this predefined number is defined as "rule updating frequency"). For example, if the rule updating frequency is 1000, the rule updating will be processed every increments or decrements of 1000 data. The comparisons of the simulation results are carried out between four methods, i.e., the proposed method with rule updating frequency of 1000, 2000 and 4000, and the clustering method of standard GNP without online rule updating.

The silhouette values obtained by the four methods are shown in Table 3, and Fig. 3. Star marks (*) on the side of silhouette values show the times when the rule updating is carried out. The rule updating frequency of 4000 shows only a slight difference from no rule updating but shows increment of silhouette values in step 5 and 9, which means that the rule updating is effectively carried out.. The best results are obtained by the rule updating frequency of 1000, where silhouette values are stable with relatively high level compared to other frequency parameters. Rule updating frequency of 2000 also shows decrements on step 3, which previous silhouette value are high enough.

5. Conclusions

This paper proposed a new rule updating mechanism for distributed database with unbalanced data growth. The simulation results of the proposed method showed the better clustering results comparing to GNP rule-based clustering without on-line adaptation. In the future, we will apply fuzzy membership functions to attribute judgment to make rules with better clustering ability.

References

1. Kaoru Shimada, Kotaro Hirasawa, and Jinglu Hu. Genetic network programming with acquisition mechanisms of association rules, *JACIII.* 10(1) (2006) 102‑111.
2. Shingo Mabu, Ci Chen, Nannan Lu, Kaoru Shimada, and Kotaro Hirasawa. An intrusion-detection model based on fuzzy class-association-rule mining using genetic network programming Systems, *Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on.* 41(1) (2011) 130‑139.
3. Mansoori, E.G., FRBC: A Fuzzy Rule-Based Clustering Algorithm, *Fuzzy Systems, IEEE Transactions.* 19(5) (2011) 960-971.
4. Sinaee, M.; Mansoori, E.G., Fuzzy Rule Based Clustering for Gene Expression Data, *Intelligent Systems Modelling & Simulation (ISMS),* 4th International Conference (2013) 7-11.
5. Zhao, JiangFei and Huang, TingLei and Pang, Fei and Liu, YuanJie, Genetic algorithm based on greedy strategy in the 0-1 knapsack problem, *Genetic and Evolutionary Computing*, 3rd International Conference on (2009) 105-107.
6. Sylvain Guinepain and Le Gruenwald. Using cluster computing to support automatic and dynamic database clustering, *Cluster Computing IEEE International Conference*, (2008) 394‑401.
7. Sylvain Guinepain and Le Gruenwald. Automatic database clustering using data mining. Database and Expert Systems Applications, DEXA IEEE 06(17) (2006) 124‑128.
8. Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures, *Data Mining (ICDM) IEEE 10th International Conference*, 10 (2010) 911‑916.