Extended Abstracts of the 2020 International Conference on Solid State Devices and Materials, VIRTUAL conference, 2020, pp773-774

A High Accuracy Cluster Validation Index Processor using Novel Global Separation and Local Dispersion based Architecture for High Efficiency Machine Learning

Hui Shen[1,4], Yitao Ma[2,3,4] and Tetsuo Endoh[1,2,3,4]
[1]Graduate School of Engineering, Tohoku Univ. Japan, [2]Center for Innovative Integrated Electronic Systems,
[3]Research Institute of Electrical Communication, Tohoku University, [4]JST-OPERA
Phone: +81-22-795-4906 E-mail: tetsuo.endoh@cies.tohoku.ac.jp

Abstract

A novel Global Separation and Local Separation (GSLD) based Cluster Validity Index (CVI) architecture is proposed to achieve accurate cluster validation under unbalanced dataset from view point of its data size and density. Moreover, with 55nm-CMOS technology, GSLD based CVI processor was designed. To improve GSLD calculation speed, the novel GSLD architecture was developed making full leverage of temporary clustering result of each iteration. The designed processor achieves 200MHz operation speed at 1.2 V, and high speed GSLD calculation of 268 clocks under data precision of 8bit. The maximum vector dimension is 2048 that is enough for large-size image features. Furthermore, high accuracy cluster validation was successfully demonstrated with the full simulation of designed processor as follows. At first, 27 texture images (resolution: 600 x 800) were correctly classified by the designed processor. Next, it is shown that calculation complexity of novel GSLD architecture is only 11.1 % of conventional VRC and general GSLD calculation. Finally, in the case of unbalanced datasets, even we applied the Manhattan distance, the designed processor shows the superiority in accuracy, nevertheless conventional VRC using the same Manhattan distance failed. From all, it is shown that the CVI processor with novel GSLD architecture is useful for future high efficiency machine learning of image processing application.

1. Introduction

Clustering is one of the fundamental methods in unsupervised machine learning which is widely used in intelligent systems for image classification such as drones, automotive vehicles. For clustering datasets with uncertain categories, many CVI methods have been developed to determine the Optimal Cluster Number (OCN) by evaluating the clustering results [1-3]. Moreover, adaptive clustering method using CVIs in each clustering trial with different cluster number was studied to automatically find the OCN with the best score. Due to high calculation burden, CVIs are hard to be implemented in hardware. Only Variance Ratio Criterion (VRC), which has been reported with superior performance of the less calculation complexity, was implemented in FPGA [4]. However, VRC has poor performance for unbalanced datasets which have large variation in cluster size and density. In our previous work, GSLD based CVI was proposed to overcome VRC's issue [5]. However, its computation complexity is not feasible for dedicated LSI design. In this paper, a hardware-friendly and high accuracy CVI processor is presented for the first time using GSLD architecture, and its superiority in accuracy and efficiency is verified with real working chip.

2. Proposed GSLD based Processor Architecture

Fig.1 illustrates the calculation components of GSLD and how it leads to accurate clustering result to unbalanced datasets in Fig.1 (a) and GSLD is defined as follows:

$$GSLD = \frac{\sum_i^c SBS_i}{\sum_i^c SWD_i} \times \frac{1}{c} = \frac{\sum_i^c n_i \times \|C_i - GG\|^2}{\sum_i^c \{\sum_j^{n_i} \|C_i - x_j\|^2 / n_i\}} \times \frac{1}{c} \quad (1).$$

GSLD is a ratio of Sum of Between-cluster Separation (SBS) with Sum of Within-cluster Dispersion (SWD) and finally normalized by cluster number c. SBS measures the inter-cluster separation using sum of squared Euclidean distance between ith cluster centroid C_i to Global Gravity (GG). In contrast, SWD measures the intra-cluster compactness using the sum of squared Euclidean distance between C_i to input data x_j ($x_j \in C_i$) which is normalized by n_i (data number of C_i). By evaluating the dispersion of clusters, GSLD can better distinguish the clusters which has differences in density or size. GSLD has its maximum value to the clustering result with OCN. Fig.2 shows the texture image clustering (k-means) result that each CVIs judges to get the best results. As shown in Fig.2, GSLD could lead to accurate clustering result compared to conventional CVIs. Fig.3 shows the architecture of proposed GSLD processor in adaptive clustering system. Considering the combination with clustering units, the proposed architecture makes leverage of temporary result of clustering units. Therefore, SBS_i in eq. (1) is calculated as follows in proposed architecture:

$$SBS_i = \sum_i^c n_i \times |C_i - GG| = \sum_i^c |SS_i - n_i \times GG| \quad (2),$$

where SS_i, n_i, GG and SWD_i is considered to get from clustering units where it always calculates the distances between x_j and C_i in each iteration of clustering. Note that, Manhattan distance which requires less hardware cost is applied to the GSLD architecture without sacrificing accuracy. Fig.4 shows its operation flow of clustering with OCN.

3. Chip Design Result and Discussion

The GSLD processor chip was designed and fabricated with 55nm CMOS technology (@200MHz, 1.2V). The chip micrograph and its specification are shown in Fig. 5. Fig.6 (a) shows the operation waveform of the processor to calculate the GSLD score for clustering of texture images in Fig.7 into 3 clusters. Fig.6 (b) shows GSLD scores for cluster number from 2 to 6 with the best score at 3-cluster, where the proposed processor complete GSLD computation for each cluster number within 268 clocks in average. Fig.8 shows the result (OCN = 3) of the proposed processor clustering 27 texture images with 600 x 800 pixels (Fig.7) into 3 clusters. As shown in Fig.9 (a), by making leverage of clustering results, the proposed architecture reduces 88.9% of complexity compared with general digital design of GSLD and VRC, when input data number N, dimension d, cluster number c is 128, 2048 and 16, respectively. More complexity reduction can be achieved with increase of N. Moreover, we compared its accuracy with squared Euclidean based VRC and Manhattan based VRC. As shown in Fig.10, even we applied the Manhattan distance in our processor, it still judges the correct number for unbalanced datasets while VRC fails.

4. Conclusions

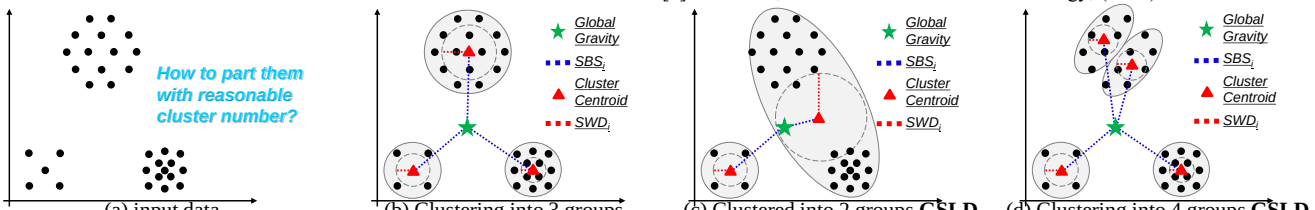
A CVI processor with novel high efficiency GSLD method is developed, which successfully realizes to decrease the complexity to 11.1% keeping notably high accuracy clustering for unbalanced datasets. The processor is especially suitable for dedicated LSI implementation of adaptive object clustering in automotive system which requires to automatically determine the OCN of unpredictable input objects.

Acknowledgements

The part of this work is supported by the Graduate Program in Spintronics, Tohoku University, JST-ACCEL Grant Number JPMJAC1301, JST-OPERA, and VDEC: The University of Tokyo with the collaboration with Synopsys, Cadence Design Systems, Mentor Corporation.

References

[1] T.Calinski et al., Commun.in Statistics-theory and Methods,3(1) (1974).
[2] X.L.Xie et al., IEEE Trans. Pattern Anal. & Mach. Intel, 13(8) (1991).
[3] S.Saitta et al., Intelligent Data Analysis, 12(6) (2008).
[4] Z.Hou et al., Jpn. J. Appl. Phys., 52 (2010).
[5] T.Li et al., IEEE Trans. Vehicular Technology, (2020).



(a) input data (unknown correct cluster number) (b) Clustering into 3 groups **GSLD score = 20** (correct) (c) Clustered into 2 groups **GSLD score = 15** (under-partition case) (d) Clustering into 4 groups **GSLD score = 10** (over-partition case)

Fig.1 Clustering of two-dimension data with different cluster number ***c***. At ***c*** = 3 the GSLD based CVI has its maximum score.

Tested Image	Expected Classification	CVIs	Previous Work	Conventional CVIs		
			GSLD [5]	**VRC** [1]	**XB** [2]	**SF** [3]
4 brick, 4 pebble, 4 rattan, 8 water	classify into **4 clusters**	Optimal Cluster	**4** clusters (correct)	5 clusters (over partition)	5 clusters (over partition)	6 clusters (over partition)
		Clustering Result Determined by CVIs				

Fig.2 Texture image classification with GSLD CVI compared with conventional CVIs.

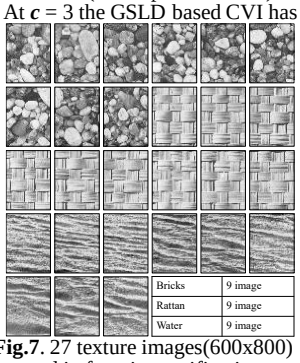


Fig.7. 27 texture images(600x800) used in function verification.

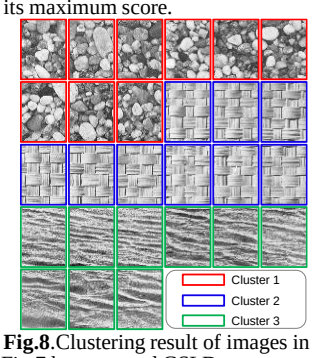
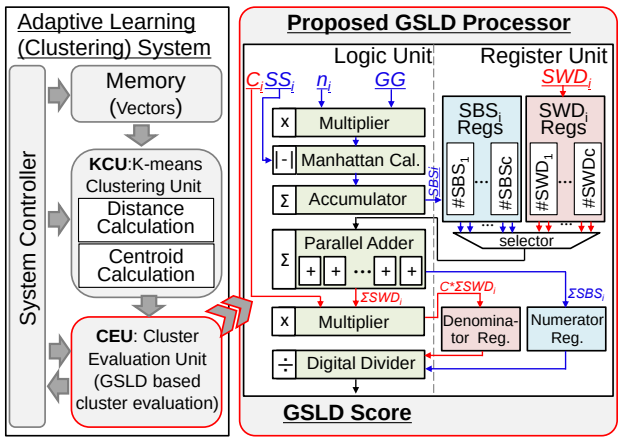


Fig.8.Clustering result of images in Fig.7 by proposed GSLD processor.



(a) GSLD based adaptive k-means system (b) Architecture of proposed GSLD processor

Fig.3. Architecture of the proposed GLSD processor in adaptive clustering (k-means) system

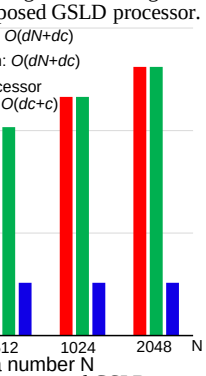


Fig.9. Reduction of computational complexity using proposed GSLD processor compared with general designed VRC and GSLD.

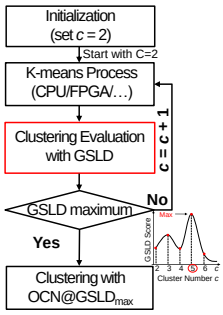
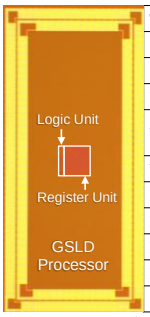


Fig.4. Operation flow



Technology	55nm CMOS
Core size	0.48*0.49 mm^2
Core Voltage	1.2 V
Frequency	200MHz
Average Computation Speed for GSLD cal.	268 clock
Vector Dimension (d)	1-2048
Number of cluster (c)	1-16
Data number (N)	1-128
Data Precision	8 bit /dimension
Distance measurement	Manhattan
Cluster Validity Index	GLSD

Fig.5. Chip photo of the proposed GSLD processor and its characteristics.

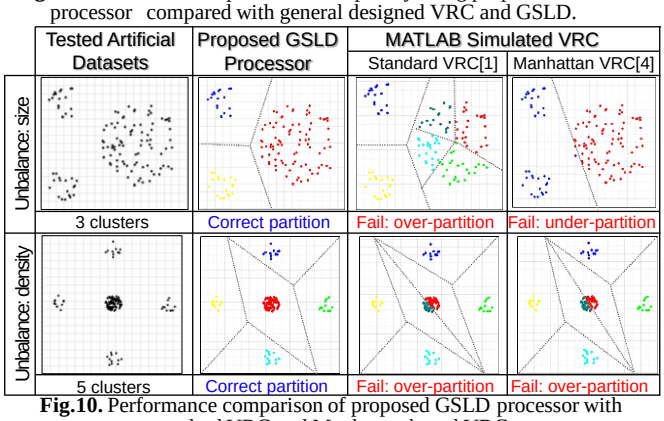
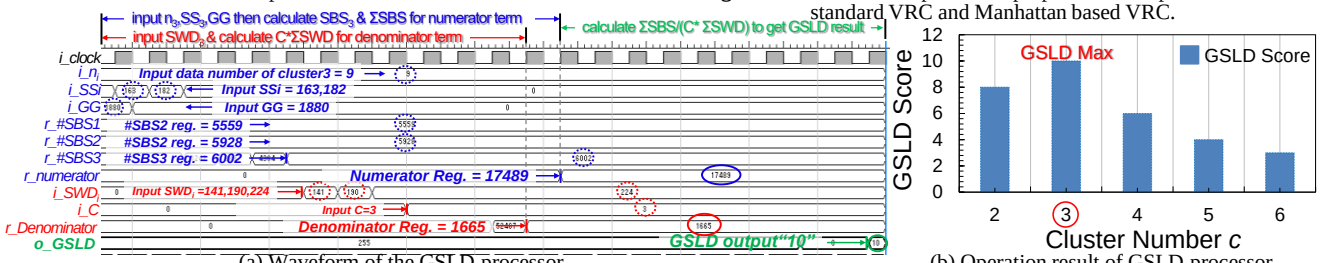


Fig.10. Performance comparison of proposed GSLD processor with standard VRC and Manhattan based VRC.



(a) Waveform of the GSLD processor

(b) Operation result of GSLD processor

Fig.6. Operation verification of the proposed GSLD processor which calculates the clustering the sample data in Fig.7.