ELectronics EXpress

IEICE Electronics Express, Vol.6, No.10, 581–586

Prostate cancer classification processor using DNA computing technique

Byung-Soo Kim[1a)], **Jae-Yeon Song, Kyu-Yeul Wang, Sang-Seol Lee, and Duck-Jin Chung**

[1] *School of Information and Communication Engineering, INHA University, 253 Yonghyun-Dong, Nam-Gu, Incheon 402–751, Korea*

a) *soc_bskim@inhaian.net*

Abstract: Recently, classification and gene selection of DNA microarray data are important in biomedical research. DNA microarray data provide useful information that can be used to discover the complex mechanism of cancer development. DNA computing techniques are alternative approaches to analyze the DNA microarray data.

In this paper, we propose the VLSI implementation of a prostate cancer classification processor. The proposed architecture uses parallel and pipelined processing to improve speed and uses cyclic random masks to reduce memory size. We evaluated the prostate cancer classification processor by testing its performance on prostate cancer microarray data. From the experimental results, the proposed architecture reduced the memory size and classification time with little loss of classification accuracy.

Keywords: DNA, VLSI, prostate cancer, microarray

Classification: Science and engineering for electronics

References

[1] B.-T. Zhang and J.-K. Kim, "DNA Hypernetworks for Information Storage and retrieval," *Lecture Notes in Computer Science (DNA12)*, 4287, pp. 298–307, 2006.

[2] J.-K. Kim and B.-T. Zhang, "Evolving Hypernetworks for Pattern Classification," *IEEE Congress on Evolutionary Computation (CES 2007)*, pp. 1856–1862, 2007.

[3] B.-T. Zhang and H.-Y. Jang, "Molecular programming: evolving genetic programs in a test tube," *Proc. 2005 Genetic and Evolutionary Computation Conf.*, vol. 2, pp. 1761–1768, 2005.

[4] C.-H. Park, S.-J. Kim, S. Kim, D.-Y. Cho and B.-T. Zhang, "Use of evolutionary hypernetworks for mining prostate cancer data," *8th International Symposium on Advanced Intelligent Systems (ISIS 2007)*, pp. 702–706, 2007.

[5] Xilinx Corp., Virtex-4 User Guide, [Online] www.xilinx.com, Jan. 2007. 5.

DOI: 10.1587/elex.6.581
Received December 01, 2008
Accepted February 12, 2009
Published May 25, 2009

ELectronics EXpress

IEICE Electronics Express, Vol.6, No.10, 581–586

1 Introduction

Recently, there has been growing interest in the potential of molecular computing that supports massive parallel processing and uses simple DNA-computing operations. A molecular evolutionary learning model and a DNA computing algorithm for exploring the potential of molecular computing are proposed by B.-T. Zhang and J.-K. Kim [1] and a hypernetworks have been presented as a probabilistic model using hypergraph structure [2]. Also, classification and gene selection of DNA microarray data are a research area of constant interest. Since microarray data offers large and complex interaction among genes, analysis of microarray data is required. Previous studies have shown that hypernetworks can be used for analyzing the microarray data and pattern classifier. Zhang and Jang [3] use a hypernetworks model to diagnose microarray data for ALL/AML leukemia and Park [4] tests the hypernetworks model on a prostate cancer problem. Software simulation of the hypernetworks model achieved a classification accuracy of 81.5%, outperforming in its best performance the standard machine learning algorithms, such as decision trees and Bayesian networks, which obtained 79.4% and 73.5% for a prostate cancer problem, respectively.

Motivated by this promising result and by the fact that the classification processes of the hypernetworks model consist of simple, parallel memory operations, we implemented the prostate cancer classification processor utilizing the massive parallelism offered by the hardware. Also, this paper presents a modified hypernetworks model for low memory hardware implementation by using the cyclic random mask method.

The rest of this paper is organized as follows. Section 2 describes the hypernetworks model and a parameter for hardware implementation. The proposed hardware architecture and the implementation of the prostate cancer classification processor are presented in Section 3. Section 4 shows the measurement results and the conclusions are presented in Section 5.

2 Modified hypernetworks model for hardware

The hypernetworks model is adapted to analyze the microarray data or pattern classification problems. First, we construct the Library $L = \{l_1, l_2, \ldots, l_m\}$ based on a set of reference patterns $R = \{r_1, r_2, \ldots, r_n\}$. Each reference pattern $r_i = \{r_{i1}, r_{i2}, \ldots, r_{ik}, y_i\}$ consists of k components (=dimensionality) in sample i and its class label y_i. L is a set of hyperedges, and $l_i = \{l_{i1}, l_{i2}, \ldots, l_{ik}, y_i\}$ is a hyperedge that contains k feature variables and a class. k, called an order, is the number of feature variables in each hyperedge. In the generation of hyperedges, every hyperedge has $1/2k \times C(n,k)$ probability, because the number of the possible hyperedges is $2k \times C(n,k)$, where n is the dimensionality of the test space of the problem and k is the order. That is, the probability distortion $P(X,Y)$ can be represented by a set of point estimators that constitute the library L of the decision list:

$$P(X,Y) \approx \frac{1}{|L|} \sum_{i=1}^{L} f_i^{(n)}(X_1, X_2, \ldots, X_n, Y) \tag{1}$$

DOI: 10.1587/elex.6.581
Received December 01, 2008
Accepted February 12, 2009
Published May 25, 2009

where $f_i^{(n)}(X_1, X_2, \ldots, X_n, Y)$ is the ith decision list order n and $|L|$ is the size of library. The hypernetworks extracts all hyperedges that are matched to the test pattern from the library. If one feature variable in the hyperedge is not equal to corresponding components of the test pattern, they are not matched. Extracted hyperedges are separated by class, and then the hypernetworks model counts the number of each separated hyperedge. Finally, the hypernetworks model classifies the test pattern as the most matched class.

To implement the hypernetworks model, we choose the parameter. The order is an important factor for the performance (accuracy) of the pattern classification. As a general rule, the lower in order the hyperedge is, the larger the matching counts of the training patterns are. In the low order, a test pattern is matched with almost all hyperedges (both the hyperedges that have a correct class and the hyperedges that have a wrong class) because hyperedges have few elements. Also, when the order is increased, a few hyperedges are matched. This tends to result in an over fitting problem. Therefore, we should choose the optimum order. In optimum order, the test pattern is more likely to match hyperedges that have a correct class than hyperedges that have a wrong class. Also, there is no over fitting problem because the test pattern matches more than a certain number of hyperdeges. Population size also affects the accuracy of the pattern classification. By increasing Population size, the approximation can be more arbitrarily accurate. There is more theoretical background can be found in [3]. In order to choose optimum order and Population size, we simulated the hypernetworks model. Figure 1 shows the classification accuracy by increasing Population size and the order. In the simulation result, we find the optimum order and appropriate Population size, 50 and 100, respectively.

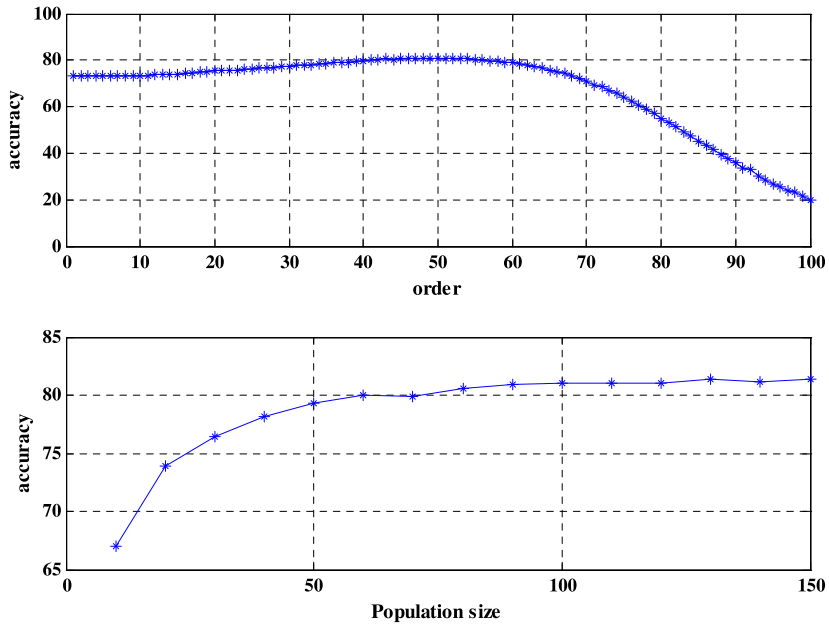

Fig. 1. Classification Accuracy as the order and Population size

DOI: 10.1587/elex.6.581
Received December 01, 2008
Accepted February 12, 2009
Published May 25, 2009

3 Implementation

The hardware architecture of the prostate cancer classification processor consists of an initialization block, a memory block, a compare block, an accumulation block, and a vote block, as shown in Figure 1.

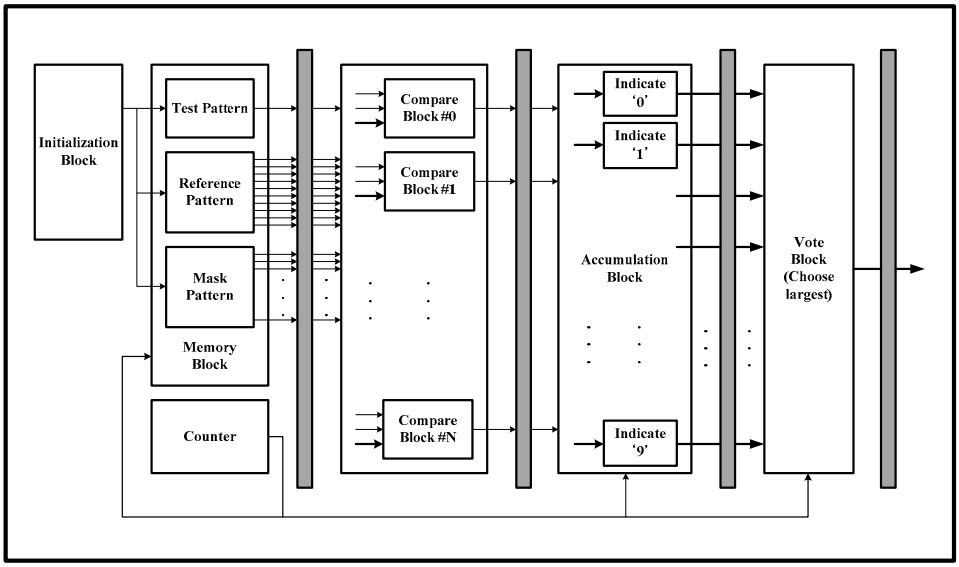

Fig. 2. Proposed hardware block diagram

In the initialization block, the set of the reference pattern and the set of the random mask are written in the reference pattern RAM and the mask pattern RAM respectively. The initialization block decides the mask pattern's selection bits, the order of the hyperedge, the Population size, and the period of the random mask set. We use the selection bits, Population size, and the period of random mask when referring to the software simulation results. In this architecture, selection bits, Population size, and the period of random mask are 50, 100, and 8, respectively.

The memory block has the test pattern RAM, the reference pattern RAM and the mask pattern RAM. The test pattern RAM and the reference pattern RAM consist of the reference pattern set (52 tumor and 50 normal samples with 225 genes) and the test pattern set (25 tumor and 9 normal samples with 225 genes). The random mask pattern RAM consists of 800 random mask patterns that have 50 ones and 176 zeros. Since the memory block uses the cyclic random mask pattern, we reduce the memory usage.

The compare block loads the reference pattern set and the random mask set from the reference pattern RAM and the mask pattern RAM. In this architecture, the compare block loads 100 random mask patterns from among 800 possible random mask patterns as well as 2 patterns–the tumor reference pattern and the normal reference pattern–from among 102 possible patterns. The compare block compares all bits in the test pattern to all bits in the reference pattern using the random mask patterns. The compare block compares the test pattern to the reference patterns when the random mask pattern's bit

DOI: 10.1587/elex.6.581
Received December 01, 2008
Accepted February 12, 2009
Published May 25, 2009

is 1. The proposed processor determines that the reference pattern is matched when 50 bits that are selected by the random mask pattern are equal to the test pattern. To determine whether the reference pattern matched or not in the compare block, a simple logical operation and the parallel processing are used.

The accumulation block adds the number of the matched patterns as a class. A signal 1 from the compare block indicates a matched pattern. This block adds 200 signals that indicate whether the reference pattern matched or not, thus it makes a critical path. To solve this problem we implement this block using multiple clocks. This block compares the matched number of each class during the two clocks.

The vote block chooses the largest number of each class, and then it generates the class of the test pattern and the number of matched patterns.

4 The result of the experiment

The proposed method is applied to classify prostate cancer data. The reference pattern set (52 tumor patterns and 50 normal patterns) and test pattern set (25 tumor patterns and 9 normal patterns) are used to verify the proposed method [4]. The proposed hardware architecture is implemented by using the Verilog HDL. We verify our architecture by using the Xilinx Virtex-4LX100 FPGA [3]. One test pattern is classified in about 1.1 us at 100 MHz using the Xilinx FPGA. The proposed hardware architecture classifies the test patterns fast, because it handles hundreds of library comparisons simultaneously using parallel processing. The proposed hardware uses the random mask method to reduce memory usage. Table I shows classification accuracy and memory usage. By using the cyclic random mask, the proposed architecture uses about 100 times less memory compared to the hypernetworks model.

Table I. Classification Accuracy and Memory Usage Comparison

Algorithm	Accuracy (%)	Memory Usage(KB)
Proposed Method	80.9	24.8
Hypernetworks	81.1	2117.5
Decision Tree	79.4	-
Bayesian Networks	73.5	-

5 Conclusion

This paper proposes hardware architecture for a Prostate Cancer classification processor. The proposed method that is based on the hypernetworks model is appropriate to implement hardware. It uses only a cyclic random mask, massive parallel processing, and a simple logical operation to classify a prostate cancer problem. We find the optimum parameter value for the proposed architecture and implement the dedicated architecture based on the hypernetworks models for Prostate Cancer classification.

DOI: 10.1587/elex.6.581
Received December 01, 2008
Accepted February 12, 2009
Published May 25, 2009

A prostate Cancer classification processor consists of an initialization block, a memory block, a compare block, an accumulation block, and a vote block. The proposed hardware architecture reduces the memory size by using a cyclic random mask, and it adapts massive parallel processing and the pipelined architecture to speed up the classification path.

It is worth emphasizing again that this implementation basically uses memory access technologies such as fetching and counting, and does not need digital signal processor (DSPs) performing precise numerical calculations, which is usually the case when solving pattern recognition problems based on conventional machine learning algorithms.

Acknowledgments

This work is supported by the Ministry of Commerce, Industry and Energy (MOCIE) through the Super Intelligence chip and applied Technology development project.

DOI: 10.1587/elex.6.581
Received December 01, 2008
Accepted February 12, 2009
Published May 25, 2009