

An Efficient Implementation of a Support Vector Machine in the FPGA

Yuki Ago, Yasuaki Ito, Koji Nakano
Department of Information Engineering
Hiroshima University

Kagamiyama 1-4-1, Higashi-Hiroshima, 739-8527, JAPAN

Abstract—The main contribution of this paper is to present an efficient implementation of a Support Vector Machine (SVM) in the FPGA. Our implementation mainly uses the DSP blocks and block RAMs in the Xilinx Virtex-6 family FPGA. Each DSP block is used to compute the product-sum performed for an internal node and the output node of the SVM. The block RAMs are used to store the weights and interim values. They also used to compute the sigmoid function. The experimental result shows that our implementation for a 128 input and 760 output nodes SVM uses 768 DSP48E1 blocks, 800 block RAMs, and 17680 slices in a Xilinx Virtex-6 FPGA XC6VLX240T-FF1156 and runs in 348.554 MHz. Also, it performs the computation for a 128 input and 760 output nodes SVM 2.72×10^6 times per second.

Keywords: Support Vector Machine, Machine Learning, SVM, FPGA, DSP48 block, Block RAM

I. はじめに

サポートベクタマシン (SVM) は人工ニューラルネットワークと呼ばれる、生物のニューロンの構造を模倣した機械学習モデルである。これらの学習モデルによって得られる識別機は画像認識のようなパターン認識処理に広く用いられている。サポートベクタマシンは教師付き学習によって2クラス間の分類を行う識別機を構成する学習モデルである。事前に用意した学習パターン (教師信号) を用いてサポートベクタマシンにどのように識別するのかを学習させる。この学習済みのサポートベクタマシンを用いて、未学習の入力データの識別を行わせる。サポートベクタマシンは多層パーセプトロン (Multi Layer Perceptron, MLP) 等の他の機械学習モデルと比較して、高い識別性能 (汎化性能) を得られるという特徴がある。これは「マージン最大化」や「カーネルトリック」と呼ばれる汎化性能を高める工夫がサポートベクタマシンの学習方法に施されているためである。

サポートベクタマシンの識別機の構造を図1に示す。サポートベクタマシンの識別機には入力層、中間層、出力層と呼ばれる3つの層があり、入力層と中間層には複数のノードが存在している。出力層のノードの個数は1個である。入力層に与えられた入力 x を中間層と出力層で処理することで識別結果 y を出力層のノードから得られる。

FPGAは内部機能を書き換えることが可能なLSIである。回路設計者はハードウェア記述言語を用いて回路データを作成しFPGAにダウンロードすることで、カスタムLSIを作成することなく所望の機能をFPGAチップ上に実装することが可能である。FPGAを利用することで、LSI内部の回路の修正したり機能を追加するといった場合でもチップ自体を交換することなく変更を加えることが出来る。その書き換え可能な特性や、FPGA自体の価格の低廉化により、通信などの規格変更が頻繁に行われる分野や教育の分野で

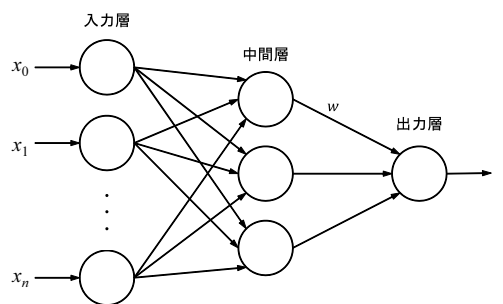


Fig. 1. サポートベクタマシンの構造

FPGAが利用されている。例えば、RSA暗号化の高速化 [1], [2] や教育向けプロセッサの開発 [3], [4]

FPGAの基本構成を図2に示す。一般的なFPGAはユーザが書き換え可能な論理ブロックであるCLBと、メモリ専用ブロックのブロックRAM(BRAM), それらを接続する内部配線, そして入出力回路から構成されている。

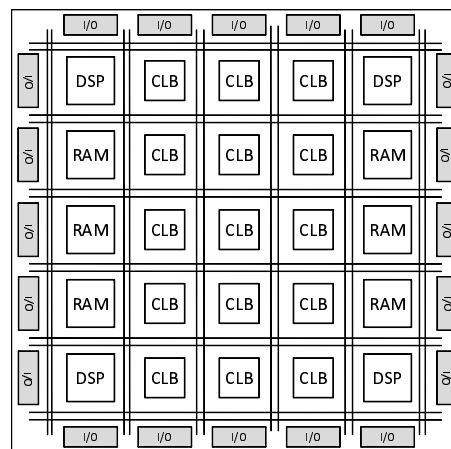


Fig. 2. FPGAの構成

さらに、本研究で使用するXilinx Virtex-6シリーズFPGAには、これらのブロックに加えてDSP48E1スライスが内蔵されている。DSP48E1スライスは25×18ビットの2の補数乗算器、48ビットの3入力加算器/累積器、48ビット論理演算器、パイプラインレジスタ、パターン検出器、Pre-Adderなどによって構成されている。

本研究ではサポートベクタマシンの識別演算回路をFPGAに実装した。識別演算とは学習済みのサポートベクタマシン

ンによって未学習のデータを処理する演算の事である。この演算回路は複数の DSP スライスを用いて、一つのサポートベクタマシン演算を実行する。DSP スライスはその機能として含まれている、パイプラインレジスタによって直列に接続されており、サポートベクタマシンの中間層の演算を実行する。この直列接続される DSP スライスの個数はサポートベクタマシンの中間層ノードの個数と等しくなっている。つまり、一つの中間層ノードで行われる演算を一つの DSP スライスが担当し、サポートベクタマシンの識別演算を並列に実行する。DSP スライスに組み込まれているパイプラインレジスタを使うことで、一定の動作周波数で任意中間層ノード数を持つ並列サポートベクタマシン演算回路を構成している。

また本研究に先立って、3 層パーセプトロン演算回路の FPGA 実装を [5] で示している。この研究では一つの 3 層パーセプトロン演算を一つの DSP と小数のブロック RAM を用いて行う実装を示したが、本研究では一つのサポートベクタマシンの識別演算に複数の DSP とブロック RAM を用いて高い並列性を持たせた実装を示す。これはサポートベクタマシンはその学習の特性として、中間層のノード数が非常に多くなるという特徴があり、複数のサポートベクタマシン演算を実装するよりも FPGA の演算資源を最大限に活用して一つのサポートベクタマシンの識別演算を高速化するためである。

II. サポートベクタマシン

サポートベクタマシンとは教師付き学習によって 2 クラスのパターン識別機を構成する手法である。サポートベクタマシンでは図 3 のように、ある入力特徴量 x を持った複数の入力データに対して、二つのクラス C_A と C_B への識別を行う。この識別は基本的には図 4 に示す線形しきい素子と呼ばれる、ニューロンを模倣した計算モデルによって行われている。

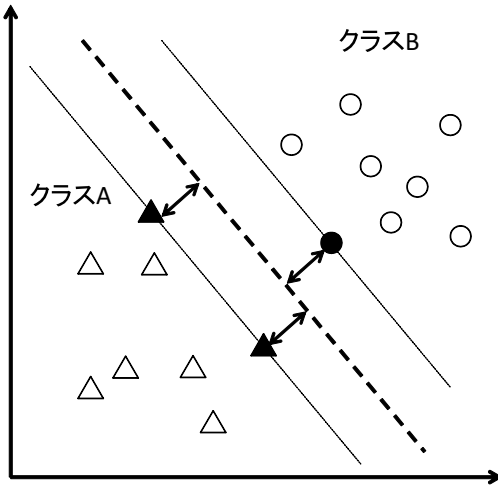


Fig. 3. サポートベクタマシンによる 2 クラス分類 (▲はクラス A のサポートベクタを表し、●はクラス B のサポートベクタを表す)

線形しきい素子は入力特徴量 x に関して、式 (1) に示す識別関数によって 2 クラス間の識別を行う。ここで w はシナプスの結合荷重に相当し、 h はしきい値である。この 2 つのパラメータ w と h を調整することで識別機を構成する。

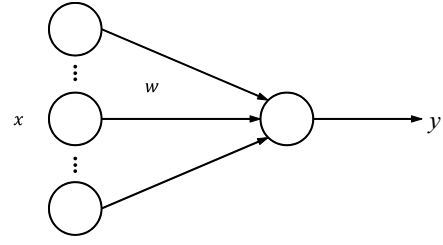


Fig. 4. 線形しきい素子

$$y = \text{sign}(w^T x - h) \quad (1)$$

$$\text{sign}(u) = \begin{cases} 1 & \text{if } u \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

サポートベクタマシンではこの線形しきい素子を構成するためのパラメータを求めるために教師付き学習を行う。教師付き学習では、まず入力特徴量 $x_i (0 \leq i < N)$ とその属するクラス C_k のラベル t_i を一つの組とした、複数のサンプルパターン (教師信号) を大量に用意する。その教師信号に対して、線形しきい素子から期待する出力が得られるようになるまでパラメータの調整を行う。サポートベクタマシンでは教師付き学習の際に、未学習の入力に対する汎化性能を向上させるために、図 3 のように、教師信号の中から識別平面の近傍に存在するものをサポートベクタとして選択し、このサポートベクタ間のマージンを最大化するようにパラメータを求める。具体的には、制約条件

$$t_i(w^T x_i - h) \geq 1 \quad (3)$$

の下で、目的関数

$$L(w) = \frac{1}{2} \|w\|^2 \quad (4)$$

を最小にする最適化問題を解くことで、二つのパラメータ w と h を求める。ここで t_i は教師信号として与えられた特徴量の属するクラス C_k に付けられているラベルである。この最適化問題を解くことで式 (5) に示す識別関数が得られる。ここで S はサポートベクタの添え字の集合である。

$$y = \text{sign}\left(\sum_{i \in S} w_i x_i^T x - h\right) \quad (5)$$

また、このようにして求められるのは線形分離可能な識別問題に対応するサポートベクタマシンの識別関数であるが、より複雑な問題に適用する為にサポートベクタマシンはカーネルトリックと呼ばれる手法を用いて非線形に拡張される。カーネルトリックでは特徴ベクトル x を非線形カーネル関数 $K(x_1, x_2)$ を用いて非線形空間に写像し、写像された先の特徴空間で線形識別を行う。カーネル関数として使われる関数にはガウスカーネル

$$K(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right) \quad (6)$$

シグモイドカーネル

$$K(x_1, x_2) = \tanh(ax_1^T x_2 - b) \quad (7)$$

などがある．この非線形に拡張されたサポートベクタマシンの識別関数を式 (8) に示す．

$$y = \text{sign}\left(\sum_{i \in S} w_i K(\mathbf{x}_i, \mathbf{x}) - h\right) \quad (8)$$

この非線形に拡張したサポートベクタマシンでは，サポートベクタと入力ベクトルをカーネル関数で計算した結果を線形しきい素子の入力特徴量として扱っていると見なすことも出来る．つまり，サポートベクタマシンの構成する識別機は図4の線形しきい素子の前段にもう一つ層を追加した図1のような構造を持つことが分かる．

III. SVM 識別演算回路アーキテクチャ

この章では本研究で提案するサポートベクタマシンの識別演算回路アーキテクチャについて説明する．サポートベクタマシンの識別関数は式 (8) で表されるが，この演算は大きく二つの要素に分割すること出来る．一つ目はカーネル関数 K の演算である．これは図1の中間層で行われる演算に相当する．カーネル関数にはガウスカーネルやシグモイドカーネルなど多くの種類があるが，本研究ではカーネル関数としてシグモイドカーネルを適用した演算回路を示す．二つ目は出力層で行われる線形しきい素子の演算である．この二種類の演算を実行するためには内積を求める必要があり，本研究ではこの内積演算を Xilinx 社 Virtex-6 FPGA に内蔵される DSP48E1 スライスを用いて実装している．またサポートベクトル，および中間層と出力層間の結合荷重，カーネル関数用ルックアップテーブルデータを保存するためのメモリとしてブロック RAM を使用している．

A. データフォーマットについて

本研究では，回路規模と計算時間を短くするために固定小数点数を使用した．本研究のシステムでは結合荷重や入力値，出力値，中間結果のデータに対して 18bit (符号ビット:1bit, 整数部:3bit, 小数部:14bit) 固定小数点数をデータフォーマットとして使用する．このデータフォーマットは S を符号ビット， I を整数部ビット， F を小数部ビットとして表すと， $SIIII.FFFFFFFF$ となる．このデータフォーマットによる離散誤差は最大で $\epsilon = 2^{-14} \approx 6.1 \times 10^{-5}$ となり，最大値は $0111.111111111111 = 8 - \epsilon$ ，最小値は $1000.000000000000 = -8$ となる．したがって，本研究の演算回路で扱うことのできる実数の範囲は $[-8, 8 - \epsilon]$ で，精度は 6.1×10^{-5} となる．入力特徴ベクトルとサポートベクタの内積演算結果の $h_i = \mathbf{x}_i^T \mathbf{x}$ は演算結果によってはこの範囲を超えることがあるが，その場合には最大値か最小値に丸めている．

B. ブロック RAM の構成

次に演算に使用する結合荷重や入力特徴ベクトルを格納するために使用するブロック RAM について説明する．本研究では格納するデータの種類に応じてブロック RAM に 3 種類の名前を付けている．

SV-RAM

サポートベクトル $\mathbf{x}_i$ の特徴ベクトルを格納するためのメモリ．パイプラインレジスタを通して伝達されて来る入力特徴ベクトルデータに合わせて，SV-RAM から対応する特徴ベクトルを順に読出し DSP に入力する．

K-RAM

カーネル関数のルックアップテーブル (LUT) として使用するメモリ．DSP スライスで求めた積和結果を入力し，このルックアップテーブルでカーネル関数の出力を求める．

W-RAM

中間層ノードと出力層ノード間の結合荷重 $\mathbf{w}$ としきい値 h を格納するためのメモリ．

これらの DSP と BRAM を使用して SVM の識別演算回路を構成する．

C. ルックアップテーブルを用いたカーネル関数の実装

本研究では SVM のカーネル関数としてシグモイドカーネル $K(\mathbf{x}_1, \mathbf{x}_2) = \tanh(a\mathbf{x}_1^T \mathbf{x}_2 - b)$ を使用する．図5にシグモイドカーネルのグラフを示す．SVM の識別演算を実装するにはこのカーネル関数の演算を実装する必要があるが，カーネル関数の演算を FPGA に乗算器などを用いて忠実に実装することは現実的ではない．そこで，本研究ではブロック RAM をシグモイドカーネルのルックアップテーブル (LUT) とすることでこれを実現している．実際には，入力の特徴ベクトルとサポートベクタの内積 $h_i = \mathbf{x}_i^T \mathbf{x}$ を DSP スライスによって求め， h_i を K-RAM にアドレスと入力する．この K-RAM にはアドレス x に $K(x)$ の演算結果が書き込まれているので， h_i を LUT のアドレスとすることでカーネル関数の演算結果が得られるようになっている．

この LUT によってカーネル関数の演算を簡略化しているが，本研究で使用するデータフォーマットである 18bit 固定小数点数の全ての演算結果を LUT に保存すると， 18×2^{18} bit の容量の BRAM が必要になる．これでは大量のブロック RAM を使用してしまうので，本研究では h_i の上位 12bit をアドレスとして使用し，LUT に使用するブロック RAM の容量を 18×2^{12} bit としている．

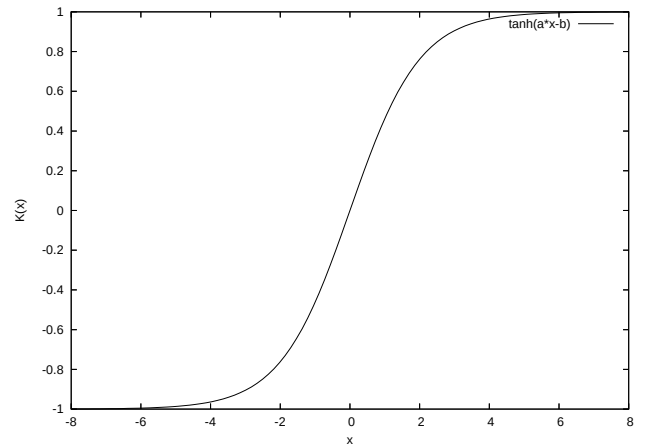


Fig. 5. シグモイドカーネル

D. DSP48E1 スライスの構成

DSP48E1 スライスには乗算器や加算器，レジスタ等が搭載されており，これらの構成をユーザが設定できるようになっている [6]．本研究では，4つの異なった構成の DSP を使用する．この 4 種類の DSP の構成を図6に示す．

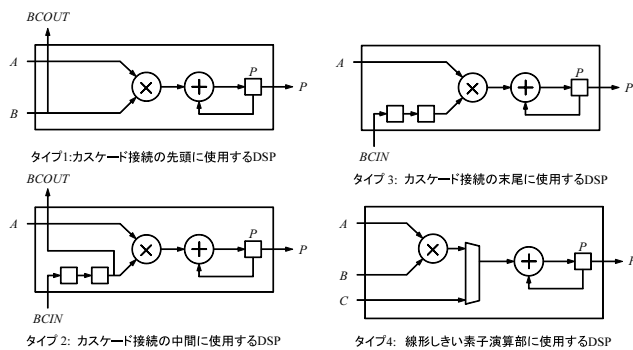


Fig. 6. 使用する 4 種類の DSP スライスの構成

この 4 種類の DSP スライスの中で、タイプ 1 とタイプ 2、タイプ 3 の DSP スライスを用いてカスケード接続した DSP スライス群を構成する。タイプ 1 の構成では 18×18 bit の乗算器と 48bit 加算器そして 48bit のレジスタ P を一つ使用する。入力 18bit の A と B、出力 18bit の BCOUT と 48bit の P となっている。タイプ 1 の構成では $A \times B + P \rightarrow P$ の演算が行われる。BCOUT は図 7 に示すように複数の DSP をカスケード接続するために使用する。

タイプ 2 の構成はタイプ 1 から 2 箇所の変更点がある。一つ目は入力 B が 18bit の BCIN となっている点で、これは直前の DSP の出力 BCOUT と接続するためである。二つ目は入力 BCIN と乗算器の間に 2 つの 18bit のパイプラインレジスタが追加されている点である。これはカスケード接続した DSP スライス間の BCOUT-BCIN 接続を通じて全ての DSP に入力データを伝達する際に、このパイプラインレジスタを通して入力データを伝達するためである。

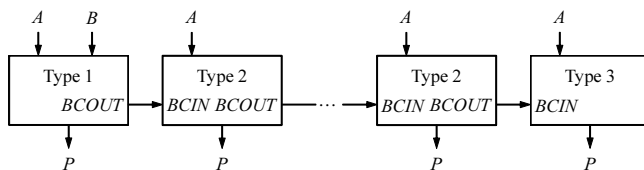


Fig. 7. カスケード接続した DSP スライス

タイプ 3 の DSP はカスケード接続の末尾に使用する DSP で、タイプ 2 と同様に 18bit の BCIN を持つが、BCOUT は持たない。これはカスケード接続に使用する BCOUT は他の DSP の BCIN に必ず接続しなければならないためである。先頭の DSP にタイプ 1、末尾の DSP にタイプ 3 を使用し、その間の DSP をタイプ 2 とすることで、先頭 DSP の B ポートへ入力された値を、次々との隣り合う DSP スライスに受け渡すことが出来る。本研究では B ポートにサポートベクタマシンの入力値を入力することで、これらのカスケード接続した DSP スライスの一つ一つが、サポートベクタマシンの一つのノードで行われる積和演算を担当するようにしている。

しかし、このカスケード接続可能な DSP スライスの個数には上限がある。本研究で使用する Vitex6 FPGA には 768 個の DSP スライスが搭載されているが、カスケード接続出

来る DSP スライスは 96 個ずつとなっている。また、カスケード接続した DSP の末尾にある DSP は BCOUT を使用できないので、複数のカスケード接続した DSP スライス群をさらに直列に接続することは出来ない。そこで、本研究では図 8 のようにサポートベクタマシンの中間層の演算を複数に分割し、分割した中間層をカスケード接続した DSP に割当てる。この複数のカスケード接続 DSP に同じ入力特徴量を入力し、その結果を最後に足し合わせることで、FPGA に搭載されている DSP を最大限に使用した演算回路を構成している。

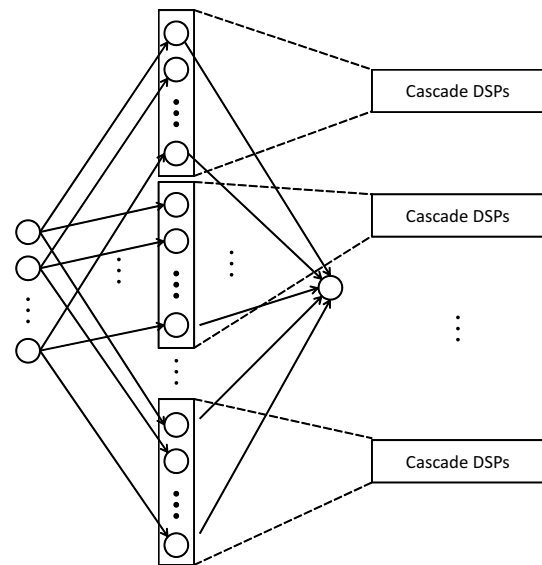


Fig. 8. SVM の中間層の分割

このカスケード接続した DSP の演算結果の加算と中間-出力層間で行われる線形しきい素子演算を実行する為にタイプ 4 の DSP を用いる。タイプ 4 の DSP は 3 つの入力 A, B, C を持ち、A には W-RAM、B にはカスケード接続 DSP の出力を接続する。C は分割された別のカスケード接続 DSP からの出力を入力するためのポートで、 $C + P$ を求めることで分割して求めたサポートベクタマシンの演算結果を一つにまとめるために使用する。つまり、カスケード接続した DSP で中間層の演算結果を求めた後、このタイプ 4 の DSP で出力層の演算を求める。

E. SVM 識別演算回路の構成

これらのカスケード接続した DSP スライスやブロック RAM を用いて構成した SVM 識別演算回路の概要図を図 9 に示す。この演算回路は大きく二つのブロックに分けることが出来る。一つはサポートベクタマシンの中間層ノードで計算される入力特徴ベクトルとサポートベクタのカーネル関数演算を担当する回路である。そして、もう一つは出力層ノードで計算される線形しきい素子演算を担当する回路である。カーネル関数演算部はカスケード接続された DSP スライスとサポートベクタの特徴ベクトルを格納する SV-RAM、そしてカーネル関数のルックアップテーブルである K-RAM から構成されている。線形しきい素子演算部は内積演算を実行するために必要な DSP 一つと線形しきい素子の結合荷重を格納する W-RAM から構成されている。

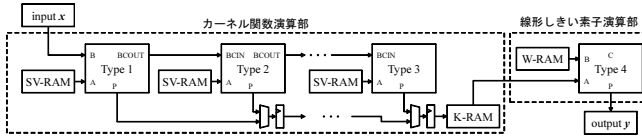


Fig. 9. SVM 識別演算回路の概要図

カーネル関数演算部のタイプ 1DSP スライスの B ポートに入力特徴ベクトル x を入力する。タイプ 1DSP に入力された入力特徴ベクトルはタイプ 1 に接続された SV-RAM から読み出されるサポートベクタ特徴量と内積が求められる。同時に BCOUT を通して次のタイプ 2DSP の BCIN に入力特徴量 x が伝達される。タイプ 1DSP に全ての入力特徴量が入力されると出力 P から一つの入力特徴量とサポートベクタ特徴量の内積 h_{DSP1} が求まる。内積 h_i は P から出力され、DSP の外側に用意されたセレクト付きパイプラインレジスタを通して K-RAM まで伝達される。タイプ 2 から出力される内積 h_{DSP2} は h_{DSP1} の直後にパイプラインレジスタに出力され、 h_{DSP1} と同様にして K-RAM に入力される。同様にしてそれ以降の DSP に関しても、入力特徴量はカスケード接続された DSP 内のパイプラインレジスタによって全ての DSP スライスに伝達され、それぞれの DSP スライスでは入力特徴ベクトルとサポートベクタの内積 $h_i = x_i^T x$ を求める。求まった内積 h_i はカーネル関数の演算結果に変換するため、DSP スライス外に用意されたパイプラインレジスタを通して K-RAM に入力される。このパイプラインレジスタには先頭のタイプ 1DSP スライスの出力結果から末尾のタイプ 3DSP スライスまでの出力結果が連なって入力される。

この内積値はカーネル関数のルックアップテーブルである K-RAM に入力される。つまり、入力特徴量と各サポートベクタ特徴量との内積が次々と K-RAM に入力され、カーネル関数で演算を行った結果がルックアップテーブルを参照することによって求められる。このように本研究では、カーネル関数演算部のパイプラインレジスタを用いてタイミングよく演算結果を伝達することで、サポートベクタマシンの中間層ノードの演算を並列に実行している。この K-RAM の出力は続く線形しきい素子演算部のタイプ 4DSP スライスに入力される。この DSP スライスでは W-RAM から読み出される結合荷重との内積 $w^T K(x_i, x) - h$ が計算され、サポートベクタマシンの演算結果がタイプ 4DSP スライスの P から出力される。

またより多くの中間層をもつサポートベクタマシンの識別演算を実行するために複数のカスケード接続した DSP を使用する回路のブロック図を図 10 に示す。図のようにタイプ 4DSP の出力を次のタイプ 4DSP に入力し、 $C + P$ を求めていくことで複数の中間層に分割して求めた出力層の演算結果を一つにまとめて出力する。

IV. 性能評価

本研究では、提案するサポートベクタマシン識別演算回路を Xilinx 社の Virtex6 FPGA 6VLX240T-FF115 を用いて実装を行った。実装結果を表 I に示す。表より、提案した回路は SVM の中間層ノード数が増えて回路規模が変化しても一定の動作周波数で動作することが分かる。またこの回路は演算結果が後の演算に影響するループバックを持たな

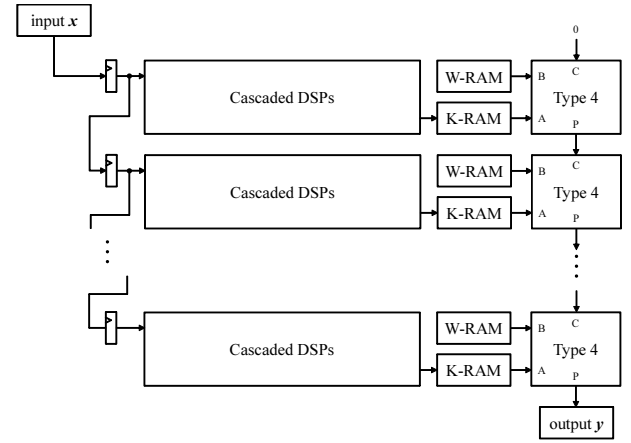


Fig. 10. 複数のカスケード接続した DSP を使用する演算回路

	[入力層ノード数]-[中間層ノード数]		
	128-95	128-475	128-760
DSP48E1 スライス	96	480	768
18k-bit ブロック RAMs	96	480	768
36k-bit ブロック RAMs	2	10	16
スライス	2210	11050	17680
動作周波数 [MHz]	348.554	348.554	348.554
レイテンシ	203	207	210
スループット [1/s]	2.72×10^6	2.72×10^6	2.72×10^6

TABLE I
SVM 演算回路の実装結果

いので、一つの入力を与えた直後に、すぐ次の入力を実行できる。したがって回路のスループットは入力データの個数にのみ依存し、回路規模が変わっても一定であることが分かる。また一つのサポートベクタマシン識別演算を実行するのに、シーケンシャルに演算を行うと、 $N_h \times (N_i + 1)$ 回の積和演算が必要となるが、本研究では中間層を分割して複数の直列 DSP で並列に演算し、更に各中間層ノードの演算をパイプライン化しているので、1つのサポートベクタマシン識別演算のレイテンシは $2N_h/P + 12 + P$ となる (N_i : 入力層ノード数, N_h : 中間層ノード数, P : 直列 DSP 群の数)。

V. まとめ

本研究では FPGA に内蔵されている DSP48E1 スライスとブロック RAM を用いたサポートベクタマシンの識別演算回路を Xilinx 社 Virtex-6 ファミリー FPGA に実装した。実装を行った演算回路の動作周波数はその回路規模に関わらず 348.554MHz となった。

REFERENCES

- [1] S. Bo, K. Kawakami, K. Nakano, and Y. Ito, "An RSA encryption hardware algorithm using a single DSP block and a single block RAM on the fpga," *International Journal of Networking and Computing*, vol. 1, no. 2, pp. 277-289, July 2011.
- [2] Y. Ito, K. Nakano, and S. Bo, "The parallel FDFM processor core approach for CRT-based RSA decryption," *International Journal of Networking and Computing*, vol. 2, no. 1, pp. 79-96, Jan. 2012.
- [3] K. Nakano, K. Kawakami, K. Shigemoto, Y. Kamada, and Y. Ito, "A tiny processing system for education and small embedded systems on the FPGAs," in *Proc. of Embedded Software Optimization*, Dec. 2008, pp. 472-479.

- [4] K. Nakano and Y. Ito, "Processor, assembler, and compiler design education using an fpga," in *Proc. of International Conference on Parallel and Distributed Systems*, Dec. 2008.
- [5] Y. Ago, A. Inoue, K. Nakano, and Y. Ito, "The parallel FDFM processor core approach for neural networks," in *Proc. of International Conference on Networking and Computing*, Dec. 2011, pp. 113–119.
- [6] X. Inc., "Virtex-6 dsp48e user guide," 2011.