AN OPTICAL NEURAL NETWORK MODEL FOR MINING FREQUENT ITEMSETS IN LARGE DATABASES

DIVYA BHATNAGAR

Jodhpur National University
Jodhpur,Rajasthan, India

A. S. SAXENA

Gwalior, India

Abstract

This paper proposes a model for mining frequent patterns in large databases by implementing Optical Neural Networks. The whole database is scanned only once and stored as a matrix. The frequent patterns are then mined from this matrix using optical neuron network. This approach is extremely efficient in data mining, as the number of database scans is effectively less than two and all the frequent patterns are computed in parallel. Appropriate techniques are designed and performed on the proposed model to achieve this efficiency.

Keywords: Optical neural network, Matrix Vector Multiplier, Frequent Patterns, Data Mining, Association Rule Mining, Large Databases.

1. Introduction

Data mining has been recognized as promising field of database research due to its wide and significant application in industry. It is a key step in the knowledge discovery process in large databases [1]. It consists of applying data analysis and discovery algorithms that under limitation of acceptable computational efficiency produce a particular enumeration of patterns over the data [2]. Due to massive amount of data generated from business transactions, there arose a need for an efficient model to discover interesting patterns from these databases in order to derive new information and knowledge for effective decision making. One of the important problems of data mining is association rules mining. A key component in association rule mining problem is to find all frequent itemsets. Many algorithms have been implemented for finding frequent patterns for data mining. In large databases the problem of mining frequent patterns gets multifold. Since the database needs to be scanned several times, efficient algorithms are required for mining frequent patterns. One of the important developments in area of association rule mining was development of Apriori [3] algorithm. It was improved by partition [4] and sampling [5], but both of these approaches were inefficient when the database was dense. PASCAL [6] is an optimization of Apriori but when all the candidate patterns are candidate key patterns, then the algorithm behaves exactly like apriori. The optical neural network model proposes the most optimized approach with only one database scan and parallel computation of frequent patterns. Traditional association rule algorithms adopt an iterative method to discovery, which requires very large calculations and a complicated transaction process [7]. This approach adopts an optical neural network method to discover frequent itemsets. The model quickly discovers frequent itemsets and effectively mine potential association rules.

2. Proposed model

The approach has been proposed for mining frequent patterns in large databases. Here, a model has been developed using an optical neural network. Optical neural networks interconnect neurons with light beams. As a result no insulation is required between signal paths. The light rays can pass through between each other without interlacing. Only the spacing of light sources, the effect of divergence and the spacing of the detectors limit the density of the transmission path. As a result all signal paths operate simultaneously. The strengths of weights are stored in holograms with high density. These weights can be modified during operation to produce a fully adaptive system. The proposed model uses electro-optical matrix multipliers where optics is used for its massive parallelism and input and output data are defined in the electronic domain.

2.1. *Neural Network*

Artificial neural networks are inspired by the operation of the human brain. It is a model of the biological neuron as a circuit component to perform computational tasks. Artificial neural networks consist of a number of simple computing elements called neurons that are modeled after the human nerve cell. Each neuron receives a number of input signals and performs a simple operation on this set of inputs. The output of each neuron is fanned out to the inputs of other neurons.[8]

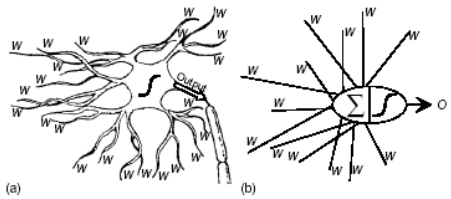
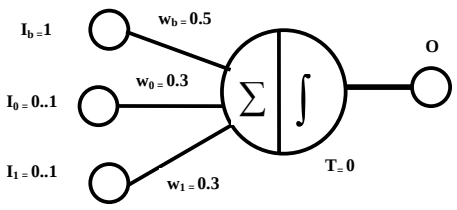


Figure 1 Human nerve cell (a) and its model (b). Weighted (*w*) input signals are added. The resulting sum is compared to a threshold as is depicted with the nonlinear, S-shaped neural response function in the cell body.[8]

In Fig. 1 a human nerve cell, or neuron, (a) and its artificial equivalent (b) are sketched. The neuron receives a set of input signals via a number of tentacles or dendrites. At the tip of each dendrite the input signal is weighted with a factor w, which can be positive or negative. All the signals from the dendrites are added in the cell body to contribute to a weighted sum of inputs of the neuron. If a weight is positive the corresponding input will have an excitatory influence on the weighted sum. With a negative weight, an input decreases the weighted sum and is inhibitory. In the cell body the weighted sum of inputs is compared to a threshold value. If the weighted sum is above this threshold, the neuron sends a signal via its output to all connected neurons. The threshold operation is essentially a nonlinear response function as is indicated in the figure with an S-shaped, sigmoid, curve. The function of a neuron can be described in mathematical form with:

$$O = F\left(\sum_i wi \,.\, Ii\right) \qquad (2.1)$$

where, O is the output signal of the neuron and Ii are the input signals to the neuron, weighted with a factor wi . F is some nonlinear function representing the threshold operation on the weighted sum of inputs.[8]



I_b	I_0	I_1	$\sum_i w_i I_i$	O
1	0	0	-0.5	0
1	0	1	-0.2	0
1	1	0	-0.2	0
1	1	1	0.1	1

Figure 2. A neural implementation of a logical AND function and the corresponding truth table including weighted sum of inputs of the neuron.

2.2. *Optical Neural Network*

Optical neural networks interconnect neurons with light beams. No insulation is required between signal paths, the light rays pass through between each other without interlacing. The density of transmission path is limited only by the spacing of light sources, the effect of divergence and the spacing of detectors. As a result all signal paths operate simultaneously, which results in a true data rate[9].

Electro-optical Matrix-Multipliers: Theses nets provide a means for performing matrix multiplication in parallel. The network speed is limited only by the available electro-optical components. The computational time is potentially in the Nanosecond range[9].

The matrix multiplier is capable of multiplying an m-element input vector by a m * n matrix, which produces n-element NET vector. The column of light sources passes its rays through a lens, such that each light illuminates a single row of weight shield. The weight shield is a photographic film in which the transmittance of each

square is proportional to the weight. There is another lens which focuses the light from each column of the shield to a corresponding photo detector. The NET is calculated by,

$$NET = \Sigma\, w_{ik}\, x_i,$$

where NET_k - NET

w_{ik} = weight from neuron i to neuron k

x_i = input vector component i.

The output of each photo detector will represent the dot product between the input vector and the weight matrix. The set of outputs is a vector equal to the product of the input vector with weight matrix. Hence matrix multiplication is done in parallel. The speed is independent of the size of the array. This makes the network to be scaled up without increasing the time required for computation. For the weights, instead of photographic film, liquid crystal light valve may be used.

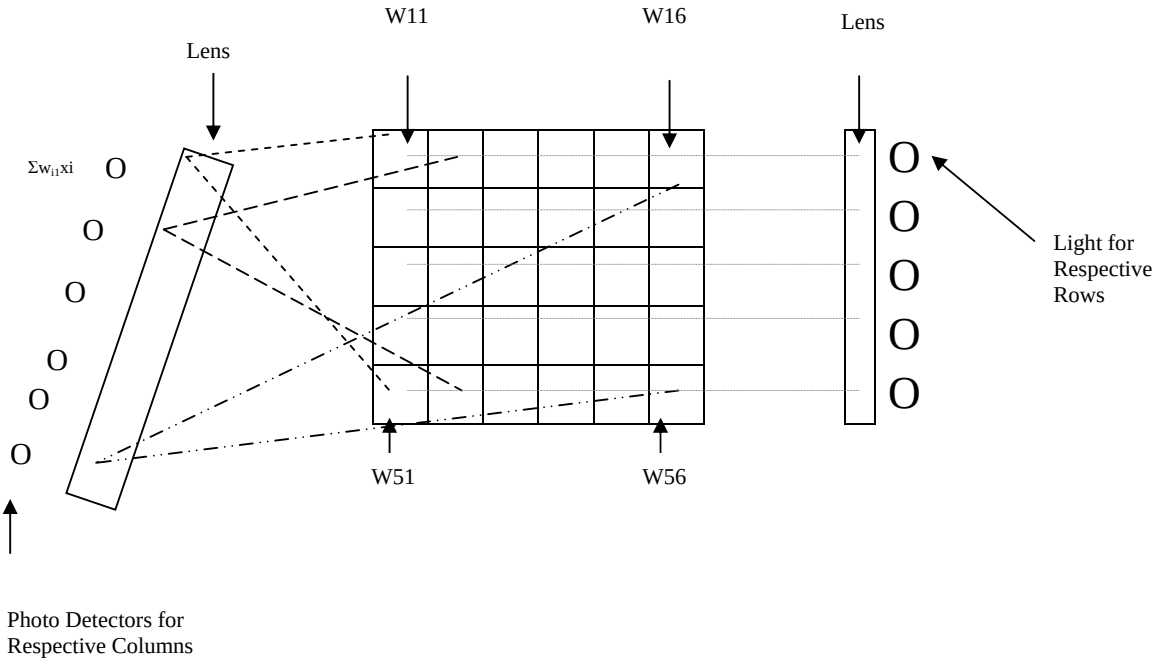


Fig. 3. Electro-optical Vector Matrix Multiplier implementing a 5 by 6 matrix.

2.3. *Mining Itemsets*

In the suggested model, each transaction is represented by rows of the weight matrix and the presence and absence of any item is stored as weights 1 and 0 respectively. A sample database D and its corresponding weight matrix M is given.

D

Row_id	Items
1	**I1 I3 I4 I6**
2	**I2 I3 I5 I6**
3	**I1 I2 I3 I5 I6**
4	**I2 I5 I6**
5	**I1 I2 I3 I5 I6**

M

I1	I2	I3	I4	I5	I6
1	0	1	1	0	1
0	1	1	0	1	1
1	1	1	0	1	1
0	1	0	0	1	1
1	1	1	0	1	1

Electro-optical implementation of the suggested model uses an array of light emitting diodes (LEDs) to represent the logic or neurons of the network. Their state will always remain on, as we need the input vector to be a unit vector consisting of all 1s. This is because we need the product of w and x to be either 1 or 0 to

represent presence or absence of the item. This is possible only if we take 1 as the input and weights as 1 or 0. The array of LEDs represents the input vector, and an array of photodiodes (PD) is used to detect the output vector. Multiplication of the input vector by the matrix M is achieved by horizontal imaging and vertical smearing of the input vector that is displayed by the LEDs on the plane of the mask M by means of an anamorphic lens system. The output obtained is the net input to the neuron i.e., NET = $\Sigma\ w_{ik}\ x_i$. It is then thresholded and compared with the minimum support required for an item to be frequent. If the NET>=min-sup, the item is frequent. In this way a single weight matrix can find all frequent-1 item sets in the database

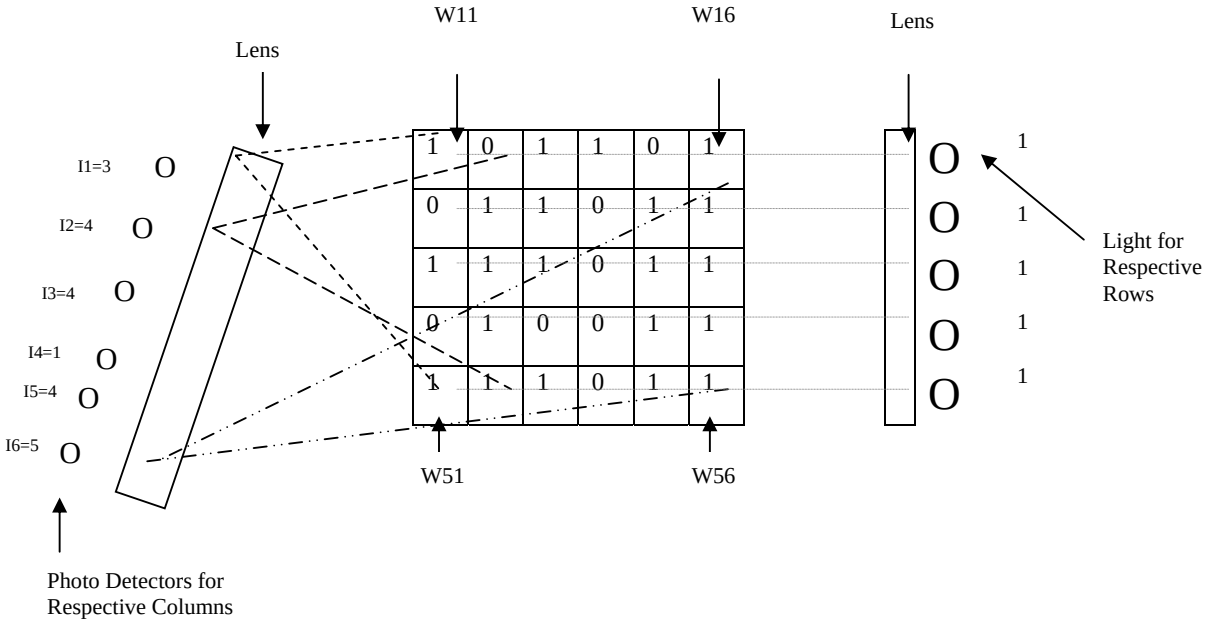


Fig. 4. Electro-optical Vector Matrix Multiplier implementing a 5 by 6 matrix with weights as 0 or 1indicating the presence or absence of an item in a transaction.

2.4. *Method of application*

The detailed method of the proposed model is presented below:

Step 1. To generate a weight matrix from transaction database

The mined transaction database is D, with D having m transactions and n items. Let $T=\{T_1,T_2,\ldots,T_m\}$ be the set of transactions and $I=\{I_1,I_2,\ldots,I_n\}$be the set of items[7]. We set up a weight matrix M_{m*n}, which has m rows and n columns. Scanning the transaction database D, if item I_j is in transaction T_i , where 1=j=n,1=i=m, the element value of M_{ij} is '1,' otherwise the value of M_{ij} is '0.' The weight matrix cell containing a 0 has a high density mask through which the light does not pass. It indicates the absence of the item. The cells containing a 1 have a transparent mask through which the light can pass easily. It indicates the presence of the item.

Step 2. To multiply the inputs with the weights

When the rays of light from the light sources fall on any column of the matrix, the light passes through all the cells containing a 1 and all these weighted sums are accumulated by the photo-detectors as the support of that itemset. This support is then stored electronically and compared with the minimum support. If the support of the item >= the min-sup, the itemset is frequent. Here, the minimum support acts as the threshold for optical neurons.

Step 3. To multiply the inputs with the weights

A second lens (which is not shown in the figure below) is used to collect the light emerging from each row of the weight mask on individual photo-sites of the PD array [10]. These are then treated as inputs for each row of the second weight mask behind it and so on. All these masks are identical copies of M. Similar process as above is followed to get the weighted sum as support.

3. Running Example

Generate a matrix M of size 5 by 6 from D as shown below. 1s are fed as inputs. The weighted sum from M1 when compared with the minimum threshold (say 3), we get frequent 1-itemsets F1 as {I1}, {I2}, {I3}, {I5}, {I6}.

Now, let M1 and M2 be the 2 frames of the matrix M. Let M1 be superimposed on M2, i.e., the output received from each row of M1is fed as input to M2. Thus, the two values of the superimposed cells get multiplied. The sum of these values is then accumulated by the photo-detectors as before and we get the support of all 2-itemsets. To get the desired results, the frame M2 should be shifted one column at a time. On shifting M2 to left by one column, we get the support count for {I1 I2}, {I2 I3},{I3 I4},{I4 I5},{I5 I6}, on shifting left by next column, we get the support count for {I1 I3}, {I2 I4}, {I3 I5}, {I4 I6}, on shifting left by next column, we get the support count for {I1 I4},{I2 I5},{I3 I6}, on shifting left by next column, we get the support count for {I1 I5}, {I2 I6}, and finally on shifting left by next column, we get the support count for {I1 I6}. On comparison with minimum support threshold the model generates {I1 I3}, {I1 I6}, {I2 I3}, {I2 I5}, {I2 I6}, {I3 I5}, {I3 I6}, and {I5 I6} as frequent 2-itemsets, i.e. F2.

The third frame generates frequent 3- itemsets, i.e., F3= {I1 I3 I6}, {I2 I3 I5}, {I2 I3 I6},{I3 I5 I6}.

The fourth frame generates frequent 4- itemsets, i.e., F4= {I2 I3 I5 I6}.

Thus, the superimposition of 2 frames gives all frequent 2-itemsets. Similarly, the superimposition of n frames can give all frequent n-itemsets. The operation shows massive parallelism and saves the computing time to large extent as compared to all other methods available.

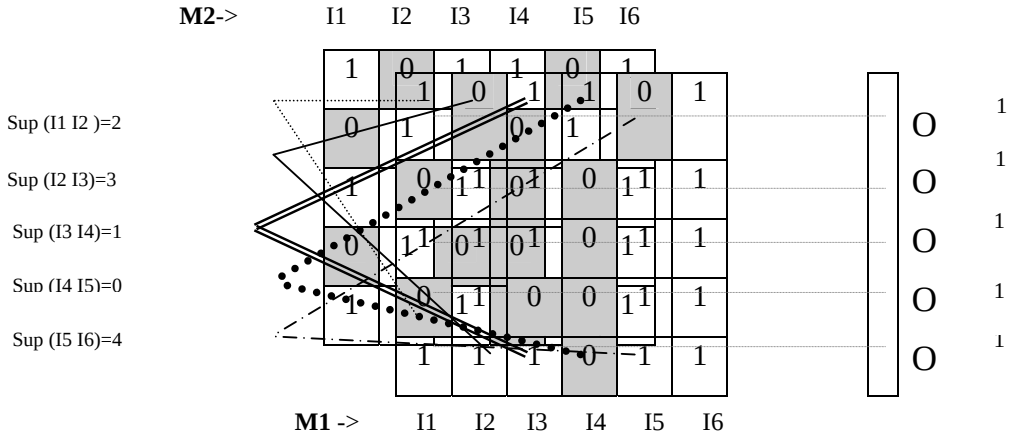


Fig. 5. Electro-optical Vector Matrix Multiplier with 2 frames M1 and M2. M2 is shifted by one column and M1 is superimposed on M2.

4. Conclusion

In this paper, a frequent itemset mining model based on optical neural networks is proposed. The main features of this model are that it only scans the transaction database once, it does not produce candidate itemsets, and it adopts the optical neural networks to discover frequent itemsets. It stores all transactions in bits, so it needs lesser memory space as compared to others and can be applied to mining large databases. The database is

accessed only once and then all the supports are determined simultaneously, thus making the process faster and more efficient than the other available techniques.

The performance of this optical model is much better than that of Apriori, PACSAL, and other popular algorithms as it does not have to generate candidate patterns. It eliminates joining and pruning making the model less time consuming. It can further be improved by replacing the electronic threshold by optical threshold. Optical threshold maintains the spatial optical parallelism and avoids opto-electronic inter-conversions [11]. The model finds its future scope in incremental data mining and online data stream mining.

References

[1] Han, J., Kamber, M.: Data mining: *Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco (2001)
[2] Fayyad U., Piatetsky-Shapiro G., Smyth P., and Uthrusamy R 9Eds.). Advances in Knowledge Discovery and Data
[3] Agrawal R. and Srikant R.: Fast algorithms for mining association rules in large databases. In Proc. 20th VLDB, pages 478-499, Sept. 1994.
[4] Sarasere A., Omiecinsky E., and Navathe S. : An efficient algorithm for mining association rules in large databases. In Proc. 21st VLDB, pages 432-444, Sept. 1995.
[5] Toivonen H. L sampling large databases for association mining rules. In Proc. 22nd VLDB, pages 134-145, Sept. 1996.
[6] Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., Lakhal, L.: Mining frequent patterns with Counting Inference, SIGKDD Explorations Vol. 2 , Issue 2
[7] Hanbing Liu and Baisheng Wang: An association rule mining algorithm based on a Boolean matrix. Data Science Journal, Volume 6, Supplement, 9 September 2007.
[8] Mos, Evert C.: Optical Neural Network based on Laser Diode Longitudinal Modes
[9] Shivanandam, S. N., Sumathi, S., Deepa, S. N.: Introduction to Neural Network using MATLAB 6.0. TATA Mc.Graw Hill.
[10] R. Ramachandran: Optoelectronic Implemenration of Neural Networks. Use of Optics in Computing. RESONANCE Sept. 1998.
[11] I. Saxena, P. Moerland, E. Fiesler, A.R. Pourzand, and N. Collings: An optical Thresholding Perceptron with Soft Optical Threshold.
[12] Pujari A. K: Data Mining: *Techniques*: Universities Press.