Mining Frequent Itemsets without Candidate Generation using Optical Neural Network

Divya Bhatnagar
Jodhpur National University
Jodhpur, Rajasthan, India

Neeru Adlakha
SVNIT
Surat, India

A. S. Saxena
SRCEM Group
Gwalior, India

ABSTRACT
We propose an efficient technique for mining frequent itemsets in large databases making use of Optical Neural Network Model. It eliminates the need to generate candidate sets and joining them for finding frequent itemsets for association rule mining. Since optical neural network performs many computations simultaneously, the time complexity is very low as compared to other data mining techniques. The data is stored in such a way that it minimizes space complexity to a large extent. This paper focuses on how this model can be helpful in generating frequent patterns for various applications. Appropriate methods are also designed that reduces the number of database scans to just one. It is fast, versatile, and adaptive. It discovers frequent patterns by using the best features of data mining, optics and neural networks.

General Terms
Frequent Itemsets, Data Mining, Association Rules, Large Databases, Candidate Generation, Patterns.

Keywords
Optical Neural Network, Weight matrix, Electro-optical Vector multiplier.

1. INTRODUCTION
Data mining 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, data mining finds its wide applications in industries. It needs efficient techniques to discover new interesting patterns very fast from these large databases in order to derive knowledge for quick and effective decision making. The problems of finding patterns are fundamental in data mining, and from the applications, fast implementations for solving the problems are needed [3]. The problem of mining frequent patterns gets multifold, in large databases since the database needs to be scanned several times. One of the important developement in area of association rule mining was development of Apriori [4] algorithm but candidate key generation remained the unsolved issue in that too. Ariori was improved by partition [5] and sampling [6], but both of these approaches were inefficient when the database was dense. Implementing optical neural network with no candidate generation and only a single database scan for mining frequent patterns is an optimized technique. Parallel computation of frequent patterns makes mining faster. Traditional association rule algorithms adopt an iterative method to discovery, which requires very large calculations and a complicated transaction process[7]. Optics offer advantages in parallelism and massive interconnectivity required in fabricating ANN [9].

2. THE PROPOSED ONN
The paper suggests for mining frequent itemsets using an optical neural network model [14].

2.1 Artificial Neural Network
Artificial neural network inspired by human brain is a model of the biological neuron as a circuit component to perform computational tasks. The function of a neuron can be described in mathematical form with: $a = f(\Sigma_i wi \cdot pi)$

where a is the output signal of the neuron and p*i* 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.

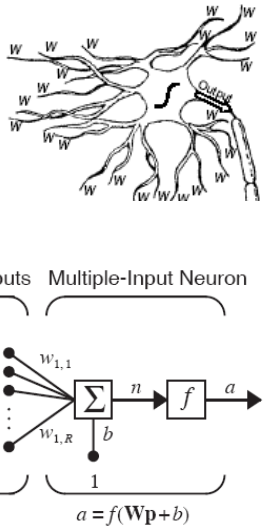

Fig. 1. Human nerve cell and Artificial Neuron Model

2.2 Optical Neural Network
In Optical neural networks the neurons are interconnected 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[8]. The strengths of weights are stored in holograms with high density. These weights can be modified during operation to produce a fully adaptive system. Weighted addition and transformation are the main operations of a neuron and since optoelectronic devices are most effective for realization of vector-matrix multiplication, the main calculation load can be put on them [20].The proposed model uses electro-optical matrix multipliers where optics is used for its massive parallelism and the input and output data are defined in the electronic domain. Electro-optical Matrix Multipliers 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[8]. The speed is independent of the size of array. This makes the network to be scaled up without increasing the time required for computation.

2.3 Network Architecture and Hardware Setup for Implementing the Proposed Model

As many k-itemsets are to be mined in parallel, multiple neurons are needed to form one layer. A single - layer feed-forward network architecture of S neurons is shown below. Each of the R inputs is connected to each of the neurons and the weight matrix has S rows [18].

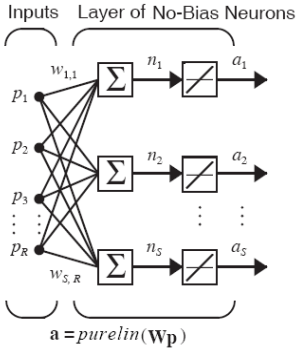

Fig. 2. A single layer feed-forward homogeneous network architecture

The layer includes the weight matrix, the summers, and transfer function boxes, and the output vector **a**. Each element of input vector p is connected to each neuron through the weight matrix W. Each neuron has a summer, a transfer function and an output a_i. Here, all neurons have the same transfer function. Such homogeneous neural networks are most suitable for optoelectronic realizations [20]. There is no bias because we do not need it in out problem. In order to find the support counts of items, the presence of an item in all transactions need to be summed up. The presence or absence of an item in a transaction is represented by 1 or 0 which is stored as weights in the weight matrix. In optics, light values can be digital binary, i.e., zero or one [19]. In our model, input vector p consists of all 1s, i.e., p = [1 1 1 1 …….1] as 1 multiplied by 1 or 0 gives the same result indicating the presence or absence of an item. Since we need the total number of 1s in all transactions, i.e., total number of 1s in each column, the transfer function suitable for our problem is linear function. Linear function gives a = n which is the support count or frequency of the itemset.

The three most basic hardware components of an optical information processing system are a source, a modulator, and a detector. Lasers are a common source of illumination as they are mathematically simpler to understand [21]. The logic functions being performed are multiplication and addition that can be implemented using Spatial Light Modulators. SLMs include photographic film, and electro-optic, magneto-optic, and acousto-optic devices. SLMs are optical masks that are controlled using electrodes [17].

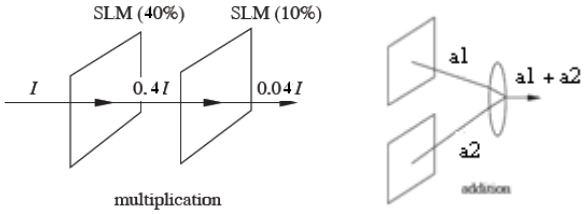

Fig**. 3. Optical implementation of multiplications and addition**

The applied voltage makes the SLM darker or reduces the amount of light projected onto an SLM. This is how multiplication takes place. Addition of optical signals is performed by reducing two light signals to one using lenses, prisms, or other devices capable of dealing with light rays. To perform a matrix vector multiplication, an SLM that is divide into R * S fields is used. The darkness in each field is kept constant to its numeric value, i.e., 1 or 0 as per the weights applied. The input vector is projected onto the SLM as shown in a sample SLM mask below.

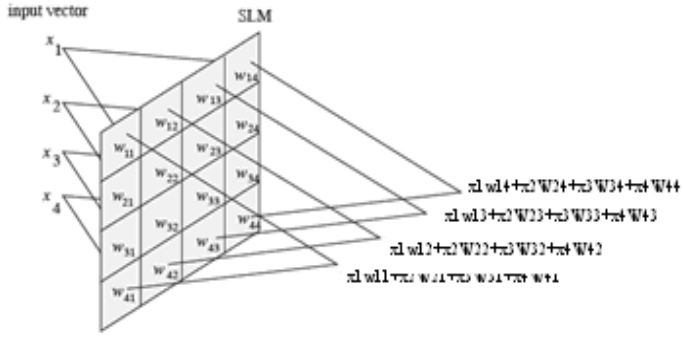

Fig. 4. Matrix-vector multiplication with SLM mask

As shown in figure 5, if the light is passed through many SLMs arranged in a stack a chain of multiplications can be computed instantly. The intensity of light received as output is proportional to the product of the darkness ration of SLMs. In order to get the support count of two or more itemsets, multiple SLMs are arranged as shown below and the NET output is obtained in the

form of outcoming light [17]. The threshold can be calculated and implemented electronically or optically [19]. Detectors most commonly used include photodiodes, highly sensitive photon detectors, and 2D detectors [21].

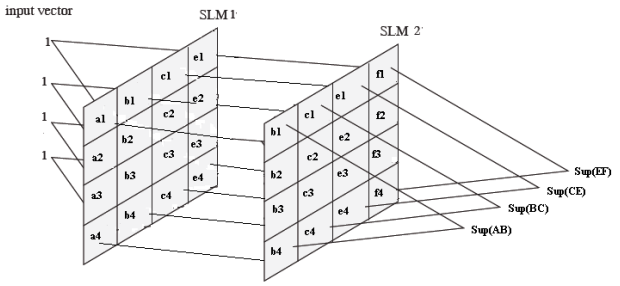

Fig. 5. Two SLMs stacked to mine 2-itemsets

3. PROCEDURE

The stepwise procedure for mining patterns is given below:

3.1 Setup the ONN

Let D be the transaction database to be mined having m transactions and n items. Let T={T1,T2,…,Tm} be the set of transactions and I={I1,I2,…,In}be the set of items[7]. Create a weight matrix Mm*n, which has m rows and n columns. Scan D database. If item Ij is in transaction Ti, the weight Wij is set to '1,' otherwise '0.' The cell of weight matrix containing a 0 has a high density mask through which the light does not pass. The cells containing a 1 have a transparent mask through which the light can pass easily. The input vector id fed as shown in Fig.4. The light falls on each row of the weight mask.

3.2 Find the support of all items.

When the light falls on the rows of the matrix, the light that passes through the cells containing the weight 1, then, 1*1 gives a 1 as the output from each cell. All these individual outputs are then accumulated by the photo-detectors as the weighted sums or the support counts for each 1-itemset.Determine all the frequent items. The support thus obtained is then stored electronically and compared with the minimum support. If the support of the item >= the min-sup, the tem is frequent.

3.3 Prune the weight matrix (optional)

If the frequency of all the itemsets is to be determined, we can continue with the initial weight mask W. This is suitable for incremental data mining approach. But if only frequent itemsets are to be found, we can do it more efficiently by removing the columns corresponding to infrequent items. This gives a new weight mask M with only frequent items.

3.4 Mine higher itemsets without candidate generation.

2 identical masks W or M as per the requirement of the problem, are taken as M1, and M2. They are stacked as shown in fig. 5 to get the weighted sum as support counts of 2-itemsets. We keep shifting the masks one column every time until all the 2-itemsets are mined. Thereafter the frequent itemsets are easily determined by comparing the frequency of itemsets with the minimum support. Similar procedure is followed to mine all the other k-itemsets by using k number of masks.

4. APPLICATIONS

4.1 Application in Association Rule Mining

Running Example: Sample database is D and its corresponding weight matrix is W [15].

D **W**

Tid	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**

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

A weight matrix W of size 5 by 6 is generated from the transaction database D. The input vector [11111] is multiplied by the weight matrix W. The output obtained is the net input to the neuron. Let minimum support be 3. The weighted sums generate the support counts for all possible 1-itemsets. When these weighted sums from W are compared with the minimum threshold, we get frequent 1-itemset F1 as {I1}, {I2}, {I3}, {I5}, and {I6}. The ONN setup is shown below:

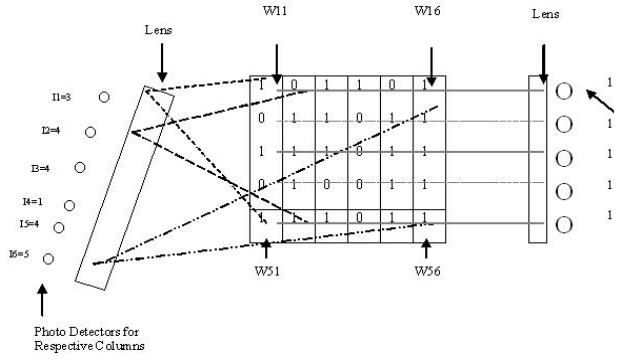

Fig. 6. ONN Model mining all 1-items.

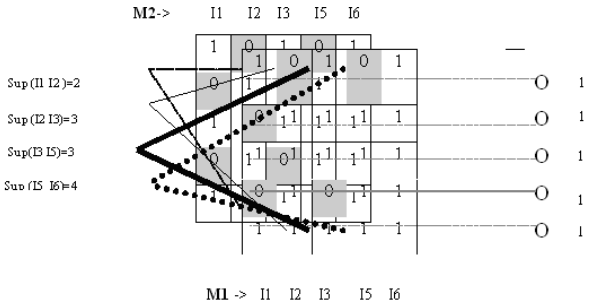

Fig. 7. ONN model mining 2 - itemsets.

Since I4 is infrequent, the new pruned weight matrix is now M. Now, let M1 and M2 be the 2 masks of M arranged as in fig. 7.

Here we get support count for {I1 I2}, {I2 I3}, {I3 I4},{I4 I5},{I5 I6}. Now M2 is composed of I2, I3, I5, and I6 to get the support count for {I1 I3}, {I2 I4}, {I3 I5}, {I4 I6}. Again M2 is shifting left by one column to get the support count for {I1 I4}, {I2 I5},{I3 I6}, on shifting M2 by one column, we get the support count for {I1 I5}, {I2 I6}, and finally on shifting M2 left by next column, we get the support count for {I1 I6}. In this way the support count of all 2-itemsets is computed. On comparison with minimum support the model generates {I1 I3}, {I1 I6}, {I2 I3}, {I2 I5}, {I2 I6}, {I3 I5}, {I3 I6}, and {I5 I6} as the list of frequent 2-itemsets, i.e. F2. Further, 3 and 4 masks generate all 3 and 4-itemsets giving F3 = {I1 I3 I6}, {I2 I3 I5}, {I2 I3 I6},{I3 I5 I6}, and F4 = {I2 I3 I5 I6} respectively.

We know that two interestingness measures are computed as:

Support (A=>B) = P(A U B) and

Confidence (A=>B) = P(B/A) = Sup (A and B) / Sup (A)

Thus for minimum confidence = 75%, some of the Strong Association Rules can be generated from above findings are:

I1 => I3 [support = 60%, confidence = 100%]

I3 => I1 [support = 60%, confidence = 75%]

I1 ^ I3 => I6 [support = 60%, confidence = 100%]

I2 => I3 ^ I5 ^ I6 [support = 60%, confidence = 75%]

4.2 Other Application Areas

Optical computing is suitable to many problems due to its Fan-in efficiency, efficiency in interconnection complexity, and energy efficiency [21]. The model presented in this paper finds scope in many areas like incremental data mining, classification, prediction, maximal approach, distributed approach, online data stream mining clustering etc.

5. CONCLUSION

In this paper, a pattern mining technique based on optical neural network model is proposed. Its main features are that it shows massive parallelism, 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. Here the database is accessed only once and then multiple supports are determined simultaneously, making the process faster and much efficient than the other available techniques. It eliminates joining since all candidate itemsets are mined automatically.

Since transfer function is applied at the end only and there is only one output layer, therefore, the model is a single layer feed-forward network. We do not need multiple layers for our problem. However, using multiple masks may be an overhead which needs to be improved by appropriate techniques. Also the input errors, weight errors, and output errors like crosstalk etc. need to be measured in order to achieve more efficiency [19].

It can further be improved by replacing the electronic threshold by optical threshold to maintain the spatial optical parallelism and avoid opto-electronic inter-conversions [10]. Since optical neural networks are not a very popular technique, the cost and the availability of resources will have to be dealt with.

The design is versatile as it can be used for various applications. However, maximal number of interconnections is a constraint. The systems with around seventy thousand interconnections have been successfully implemented [16] which is encouraging for designing such models. Some data compression techniques can further be adopted to reduce the size of the weight masks to cope up with the limitation of interconnections. The model can be optimized using maximal itemset approach, distributing the database, and applying other appropriate methods to it.

6. REFERENCES

[1] Han, J., Kamber, M. 2001: Data mining: *Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco.

[2] Fayyad U.,Piatetsky-Shapiro G., Smyth P., and Uthrusamy R 9Eds.) 1996: Advances in Knowledge Discovery and Data Mining. AAAI press, Menlo Park, CA, ISBN:0-262-56097-6, pages:611.

[3] Takeaki Uno, Masashi Kiyomi, Hiroki Arimura: LCM ver 2. Efficient mining algorithms for Frequent/ closed/ maximal itemsets.

[4] Agrawal R. and Srikant R. 1994: Fast algorithms for mining association rules in large databases. In Proc. 20th VLDB, pages 478-499.

[5] Sarasere A.,Omiecinsky E., and Navathe S. 1995 : An efficient algorithm for mining association rules in large databases. In Proc. 21st VLDB, pages 432-444.

[6] Toivonen H. L 1996: Sampling large databases for association mining rules. In Proc. 22nd VLDB, pages 134-145.

[7] Hanbing Liu and Baisheng Wang 2007: An association rule mining algorithm based on a Boolean matrix. Data Science Journal, Volume 6, Supplement, 9.

[8] Shivanandam, S. N., Sumathi, S., Deepa, S. N.: "Introduction to Neural Network using MATLAB 6.0. TATA Mc.Graw Hill."

[9] R. Ramachandran 1998: Optoelectronic Implemenration of Neural Networks. Use of Optics in Computing. RESONANCE.

[10] I. Saxena, P. Moerland, E. Fiesler, A.R. Pourzand, and N. Collings: An optical Thresholding Perceptron with Soft Optical Threshold.

[11] Tenanbaum M., Augenstein M.J., Langsam Y.: "Data Structures using C and C++ : Prentice-Hall India, Edition 2."

[12] Pujari A.K.: "Data Mining:*Techniques*: Universities Press."

[13] Mos, Evert C.: Optical Neural Network based on Laser Diode Longitudinal Modes.

[14] D. Bhatnagar, K.R. Pardasani. 2008: Mining Patterns Using Optical Neural networks: In Proc. International Conference on BUSINESS DATA MINING.

[15] D. Bhatnagar, A. K. Saxena 2011: An Optical Neural Network Model for mining Frequent Itemsets in Large Databases, Indian Journal of Computer Science and Engineering, Vol 2, No. 2, pages 212-217.

[16] I. Saxena, E. Fiesler: An adaptive Multilayer Optical Neural Network Design, IDIAP TR-94-04.

[17] R. Rojas 1996: Neural Networks. Springer-Verlag, Berlin.

[18] M. T. Hagan, H. B. Demuth, M. H. Beale: Neural Network Design

[19] E. Horward, Michel, Abdul A. S. Awwal: Analysis and Evaluation of Electro-Optic Artificial Neural Network Performance in the Presence of Non-Ideal Components.

[20] Nikolay N. Evtihiev, Rostislav S. Starikov, Boris N. Onyky, Vadim V. perepelitsa, Igor B. Scherbakov 1994: Experimetal investigation of the performance of the optical two-layer neural network, SPIE Vol. 2430 Optical Neural Networks, pages 189-197.

[21] Damien Woods, Thomas J. Naughton 2009: Optical Computing.