International Journal of Soft Computing 2 (3): 378-381, 2007

Application of Modified Simulated Annealing to the Biclustering of Gene Expression Data

[1]S. Jayalakshmi and [2]S.P. Rajagopalan
[1]Department of Computer Science, S.D.N.B.Vaishnav College For Women, Chromepet, Chennai-600 044, India
[2]Mohamed Sathak Group of Educational Institiutions, Chennai-600 005, India

Abstract: In a gene expression data matrix a bicluster is a submatrix of genes and conditions that exhibits a high correlation of expression activity across both rows and columns. The problem of locating the most significant bicluster has been shown to be NP-complete. Heuristic approaches such as Cheng and Church's greedy node deletion algorithm have been previously employed. It is to be expected that stochastic search techniques such as evolutionary algorithms or simulated annealing might improve upon such greedy techniques. In this study we show that an approach based on modified simulated annealing is well suited to this problem and we present a comparative evaluation of simulated annealing and node deletion. We show that modified simulated annealing discovers more significant biclusters.

Key words: Biclustering, gene expression data, modified simulated annealing

INTRODUCTION

DNA microarray technologies has revolutionised gene expression analysis and facilitated to monitor the expression of thousands of genes in parallel over many experimental conditions (e.g., different patients, tissue types and growth environments), all within a single experiment Lander (1999). The results from these experiments are usually presented in the form of a data matrix in which rows represent genes and columns represent conditions. Each entry in the matrix is a measure of the expression level of a particular gene under a specific condition. Thorough analysis of these datasets aids in the annotation of genes of unknown function and the discovery of functional relationships between genes. This ultimately contributes to the elucidation of biological systems at a molecular level (Berrer *et al*., 2003). Gene expression datasets typically contain thousands of genes and hundreds of conditions and mining functional and class information from such large volumes of data presents a far from trivial task. One of the main methods used thus far to investigate the underlying structure of gene expression datasets has been cluster analysis. In this approach genes showing similar expression activity over the set of conditions are grouped together into clusters. The premise behind this is that similarly behaving genes may be co-regulated and share a related function i.e., belong to a common pathway or a cellular structure. Conditions too may be clustered enabling disease types such as cancers to be defined in terms of their unique expression profiles (Pomeroy *et al*., 2002). Gene expression datasets are continually growing in size as more experiments are carried out and as experimental capacity improves. As datasets increase size it becomes less likely that objects (genes) will retain similarity across all attributes (conditions) making clustering problematic. Furthermore it is not uncommon for the expression of genes to be highly similar under one set of conditions and yet independent under another set (Ben-Dor *et al*., 2003). Clustering genes over a subset of similar conditions would be more beneficial in such cases. This approach has been termed biclustering and was first introduced to gene expression analysis by Cheng and Church (2000). Greedy search algorithms start with an initial solution and find a locally optimal solution by successive transformations that improve some fitness function. Stochastic methods such as Simulated Annealing (SA) (Kirkpatrick *et al*., 1983) improve on greedy search due to their having the potential to escape local optima. In this study we present a biclustering technique based on Modified Simulated Annealing (MSA) that improves on results and we carry out a comparative evaluation using real gene expression dataset and show that our MSA based approach finds more significant biclusters in yeast dataset.

Corresponding Author: S.Jayalakshmi, Department of Computer Science, S.D.N.B.Vaishnav College For Women, Chromepet, Chennai-600 044, India

BICLUSTERING

Biclustering refers to the 'simultaneous clustering' of both rows and columns of a data matrix (Mirkin, 1996). Cheng and Church defined a bicluster to be a subset of genes and a subset conditions with a high similarity score, where similarity is a measure of the coherence of genes and conditions in the subset. A group of genes are said to be coherent if their level of expression reacts in parallel or correlates across a set of conditions. Similarly, a set of conditions may also have coherent levels of expression across a set of genes. Cheng and Church developed a measure, called the mean squared residue score, which takes into account both row and column correlations and therefore makes it possible to simultaneously evaluate the coherence of rows and columns within a sub-matrix. They thus defined a bicluster to be a submatrix composed of subsets of genes and conditions with a low mean squared residue score (the lower the score the better the correlation of the rows and columns). The residue score of an entry aij in a bicluster B (IJ) (where I is the subset of rows and J the subset of columns making up the bicluster) is a measure of how well the entry fits into that bicluster. It is defined to be:

$$R(a_{ij}) = a_{ij} - a_{Ij} - a_{iJ} + a_{IJ} \quad (1)$$

where aiJ is the mean of the ith row in the bicluster, aI j is the mean of the jth column and aIJ mean of the whole bicluster. The overall mean squared residue score is:

$$H(I, J) = (1 / |I||J|) \sum_{i \in I, j \in J} R(a_{ij})^2 \quad (2)$$

The next problem to be tackled is how locate these low scoring biclusters within a parent data matrix. The deterministic approach is to sequentially run through all the possible combinations of rows and columns of the data matrix and find the sub-matrices which satisfy a predefined low score, δ(the set of δ-biclusters). The most significant biclusters, the largest δ-biclusters, would be of most interest as they capture the relationships between the largest number of objects. However the number of possible sub-matrices increases exponentially with the size of the parent matrix making this task practically impossible when the matrix exceeds the fairly modest size of a few 100 elements. Cheng and Church designed a set of heuristic algorithms to locate these δ-biclusters sequentially in a top-down manner by deleting the row and column nodes from the parent matrix which most improve the mean squared residue score. Upon reaching the δ threshold a node addition phase is then carried out to add rows/columns which may have been missed. Inversely correlated rows, which may represent negatively regulated genes, are also added at this stage. A subsequent study noted that as with other greedy searches there is a possibility that the system may become trapped at a locally good solution. It is thus unlikely that the global maximum or maximal δ-bicluster will be found. Applying a stochastic search technique to locate this global maximum seems to be the next logical step in the bicluster search problem.

SIMULATED ANNEALING

Simulated annealing is a well established stochastic technique originally developed to model the natural process of crystallisation and later adopted to solve optimization problems.

As with a greedy search it accepts all changes that lead to improvements in the fitness of a solution. Evolutionary optimization schemes employing the mean squared residue function have been used to tackle the bicluster search problem (Aguilar-Ruiz and Divina, 2005). These attempts failed to find more significant solutions than the Cheng and Church technique in terms of bicluster size and instead focused on returning sets of smaller biclusters with high row variability. In the virtual environment the temperature of the system is lowered after certain predefined number of accepted changes, successes, or total changes, attempts, depending on which is reached first. The rate at which temperature decreases depends on the cooling schedule. Simulated Annealing has been applied to such problems as the well known travelling salesman problem (Binder and Stauffer, 1985) and optimisation of wiring on computer chips (Kirkpatrick *et al*., 1983) and recently to biclustering of gene expression data.

EXPERIMENTAL METHODS

In biclustering using simulated annealing several parameters are common to every simulated annealing implementation.The mean squared residue score was used as a measure of bicluster fitness in this study. Many simplified cooling schedules have been introduced for practical problem solving and a popular simple cooling model is T(k) = T(k-1) / (1+ σ). Consequently each subsequent temperature is reduced. In Simulated Annealing it is also important to ensure that an adequate search is performed at each temperature. This is dictated by the number of attempts that occur before each reduction in system temperature. The selection of the

number of successes and attempts depends on the depth and size of the search space as determined by the size and dimensionality of the dataset. Our Modified Simulated Annealing Biclustering (MSAB) algorithm begins the search in a top-down manner with the initial solution containing all rows and columns. The solution is then iteratively perturbed by the deletion or addition of rows or columns with the mean squared residue being recalculated each time. The method for generating a new solution is explained below.(MSAB algorithm) This method takes into account the number of rows and columns in the current solution and a minimum solution size of 10×10 was chosen. This was deemed to represent the minimum significant size of a solution in this study. So for example, if genes correlate over 10 conditions it is likely that they may be related. This minimum solution size also prevents the search from ending on a trivial bicluster of one row or one column and score 0. To allow the comparison of MSAB with the node deletion algorithm, some way needed to be found to return biclusters of a chosen δ value. Upon reaching a δ-bicluster the minimum solution size is then reset to that of the δ-bicluster.The Modified Simulated Annealing also continues but with the added proviso of accepting solutions less than or equal to the δ- score that are larger in size. This gradually increases the size of the δ-bicluster.

Modified simulated annealing biclustering algorithm:

MSAB(f, x0, t0, rate, a, M, l, DELTA)
f: Fitness function, x0: Initial solution, t0: Initial temperature, rate: Temperature fall rate, a: Attempts, M: Datamatrix, l: Minimum solution size threshold, DELTA: Mean squared residue threshold.

```
{
t=t0
rowx=no. of rows in M
colx=no. of columns in M

while(t>tmin)
{  account=0;
colarr[]=generateColComb, range (0,colx)
newrow=generateRandomRow, range (0,rowx)
xRows[]=newRow
while(acount<attempts)
{
x0=biCluster(newRow,noOfRowSel)
while (true)
{
row=generateRandomRow, range (0,rowx)
if(row!=xRows[])
{
newRow=row
xRows[]=newRow
exit loop
}
}
acount++;
t=cool(t,rate)
xRows[]=null;
}
}
biCluster(newRow, noOfRowsel)
{
r=newRow
addNewrRow(r)  to generate Xnew
atm[]=r
loop=0
while(loop<noOfRowsel)
{
while(TRUE)
{
row=generateRandomRow, range(0,rowx)
if(row!=atm[])
{
r=row
atm[]=r;
exit loop
}
}
if (fitfunc (Xnew)<=DELTA)
addNewRow (r) to generate Xnew
else
deleteOldRow () to generate Xnew
loop=loop+1
}
return xNew
}
```

Upon the discovery of a bicluster Cheng and Church masked the solution with randomly imputed numbers from the same range as the dataset. This prevents the bicluster from being rediscovered by the deterministic node deletion algorithm. Typically, using the parameters given above and for a dataset of 2884 genes and 17 conditions the search produced large number of biclusters and much reduction in time is also observed. The generated biclusters are checked for repetition and if there is repetition they are ignored and only new clusters are discovered for solution. Cheng and Church chose a yeast cell cycle dataset in their study. This dataset contains 2,884 genes and 17 conditions and the same has been used for this study.

Table 1: Comparison of biclusters discovered in real dataset (Yeast data)

	ND	ND2	SAB	MSAB
δ		Biclusters		
300	15165	15750	16460	24500
200	8463	9540	10360	14570
100	2520	2700	2940	3740

EVALUATION OF BICLUSTERING USING MODIFIED SIMULATED ANNEALING

Cheng and Church carried out node deletion on the yeast dataset mentioned above and used a mean squared residue threshold (δ) of 300 Eq. 2. The MSAB algorithm was applied to the same yeast dataset. In this study δ thresholds of 300, 200 and 100 were set and the size of the discovered biclusters compared. MSAB produces biclusters of at least 10 columns (conditions) in width. The size of the biclusters found by ND (Node Deletion), ND2 (Adjusted Node Deletion), SAB(Simulated Annealing Biclustering) and MSAB over the various δ thresholds for the yeast dataset. The results for all δ scores are shown in Table 1.

CONCLUSION

It has been shown in the previous section that MSAB has the ability to retrieve more significant biclusters.Using MSAB we have shown that stochastic methods have the potential to give improved results for the bicluster search problem. MSAB also works in top-down manner with the mean squared residue function promoting the deletion of rows/columns which do not fit in with the trends in the dataset and takes lesser time to converge on a bicluster solution.

REFERENCES

Aguilar-Ruiz, J.S. and F. Divina, 2005. Evolutionary computation for biclustering of gene expression. In Proceed. ACM sym. Applied Computing, ACM Press. pp: 959-960.

Ben-Dor, A.., B. Chor, R. Karp and Z. Yakini, 2003. Discovering local structure in gene expression data: the order-preserving submatrix problem. J. Comput. Biol., 10: 373-84.

Berrer, D., W. Dubitzky and S. Draghici, 2003. A Practical Approach to Microarray Data Analysis, chapter 1, Kluwer Academic Publishers, pp: 15-19.

Binder, K. and D. Stauffer, 1985. A simple introduction to Monte Carlo simulations and some specialized topics, chapter Applications of the Monte Carlo Method in Statistical Physics, Spring-Verlag, Berlin, pp: 1-36.

Bleuler, S., A. Preli´c and E. Zitzler, 2004. An EA framework for biclustering of gene expression data. In Congress on Evolutionary Computation (CEC-2004), Piscataway, NJ, IEEE., pp: 166-173.

Cheng, Y. and G.M. Church, 2000. Biclustering of expression data. In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biol. ISMB., pp: 93-103.

Dhillon, I.S. S. Mallela and D.S. Modha, 2003. Information theoretic coclustering. In: Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,.

Kirkpatrick, S., C.D. Gelatt and M.P. Vecchi, 1983. Optimization by Simulated Annealing. Sci., 220: 671-680.

Lander, E.S., 1999. Array of hope. Nat. Genet., 21: 3-4.

Metropolis, N., Rosenbluth, Rosenbluth, Teller and E. Teller, 1958. Equations of state calculations by fast computing machines. J. Chem. Phys., 21: 1087-1092.

Mirkin, B., 1996. Mathematical Classification and Clustering. Dordrecht: Kluwer.

Pomeroy, S.L., P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y. Kim, L.C. Goumnerova, P. Black, C. Lau, J.C. Allen, D. Zagzag, J.M.O. and T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, D. Stolovitzky, D. Louis, J. Mesirov, E. Lander and T. Golub, 2002. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature, 24: 436-42.

Tanay, A., R. Sharan and R. Shamir, 2002. Discovering statistically significant biclusters in gene expression data. Bioinformatics, 18: 36-44.

Tavazoie, S., J.D. Hughes, M. Campbell, R.J. Cho and G.M. Church, 1999. Systematic determination of genetic network architecture. Proceedings of the National Academy of Sci., USA., 22: 281-285.