DOI Number: 10.37506/v14/i1/2020/ijfmt/193089

Associations System for Breast Cancer Microarray Data

Kain R. Qasim

University of Information Technology and Communications, Iraq

Abstract

The aim of this paper is to give biologists a tool to explore reasons and impacts of the breast cancer as patients with the same stage of illness can have different treatment responses. This paper proposes a Breast Cancer Associations system (BCA) to discover and interpret the associations among the breast cancer patient's gene expressions data. The data used in this paper is the array data of 24.483 gene expression measurements recorded for 19 breast cancer patients. BCA consists of: data preprocessing, and data mining. In the first process in BCA, the data is carried out four preprocessing steps to be suitable and enhance the second process in BCA. These four steps are data filtration, normalization, discretization, and data adaptation. The mining process stage uses a new algorithm called Row Intersection Support Starting (RISS), which traverse the row enumeration space using the user-defined mines up threshold as a starting point deploying a new data format called Row Set (RS). The last stage in the system concerns the production of the association rules based on the user defined minimum confidence threshold. Fifteen different experiments have been conducted with different parameters. The results of the experiments are recorded and compared

Keywords: *Breast cancer association's data, microarray data, data adaptation and Row Intersection Support Starting.*

Introduction

Gene expression dataset provides information on the variation of gene activity across conditions. It contains numbers which characterize the expression level of a particular sample [1]. It allows the identification of genes that are differentially expressed linked to given condition which help in understanding the response to treatments as well as provide deep insight into the nature of many diseases and lead in the development of new drugs[1-3] For example ,it indicates which genes will be over expressed and which will be under expressed when a breast cancer treatment is given to the cells .Since the number of samples is usually limited and the number of genes is measured in thousands ,such dataset are characterized as very high dimensional ones [4] .Association rule mining is the task of finding useful correlation between items in a dataset .They are rules of the form LHS -↔-RHS ,where LHS and RHS are item sets and LHS ∩ RHS= ⊘ .A brief survey for association rule mining algorithms found in [5,6]. While frequent Item set are all item sets whose support is at least equal to the minus threshold ,closed frequent item set is a condensed representation of frequent item sets .A frequent item set X is closed if there is no superset Y such that Y ⊃ X with sup (X) = sup(Y). Since the set of all frequent item sets and their exact support can be extracted from the closed frequent item set ,new researches mines closed frequent item sets [6]. The very high dimensional data like the gene expression needs special data mining techniques to discover closed frequent item sets. IRG are a set of rules that are generated from the same group of rows and meet user interestingness constrains including minus ,mincing ,and minimum chi-square (Minch) threshold. Mine Top-K focuses on the same problem of FARMER but generates most significant Top-K covering rule groups rather than generating IRG. Top-K covering rule groups are defined by first setting criterion for ranking and applying it to the resulted rules in the dataset. All of these works performed their experiments on the same data with different minus and different row lengths [11] . In most cases, COBBLER and FARMER outperform CARPENTER. The second category algorithms are IIMA, charm , and RERII [15-17]. IIMA mines all frequent item sets separately in parallel mode and then joins the results which are returned in a compressed Set Enumeration (SE) and finally generates the rules.

BCA System

BCA is a proposed system that discovers the associations among breast cancer patients' data . BCA consists of data preprocessing and data mining process, Fig. l.

Data Preprocessing

a) Filtration

Data filtration task is done in two steps according to the recommendations of a domain specific specialist. The first step is removing redundancy in the data while the second step is a data domain requirement that filters data according to a threshold. Biologists specify a certain threshold to get closer to certain data values which give more insight on the disease under study . Genes expression values greater than the threshold are excluded from the mining process.

b) Normalization

The dataset is normalized using a Mat Lab Built in function called manorm. Manorm scales the values in each column by dividing it by the column mean.

C) Discretization

The dataset is discretized using the mid-range-based cutoff method. The mid-range value for each gene is the mean value for its corresponding column. Values below or equal to its mid-range are set to 0 otherwise is set to 1. Geneexpressions with zero value indicate under expressed while those with value one indicates over expressed.

d) Adaptation

In this task, each gene is presented in two columns; one columnis for the over expressed state of the gene and the other is for the under expressed state. This adaptation is used to· produce rules that specify the state of the genes in addition to their associations. For example, the rule GI + → G2 — implies that there is an association between GI and G2, when GI is over expressed, G2 becomes under expressed.

Data Mining Process

a) Closed Frequent Item sets Generation

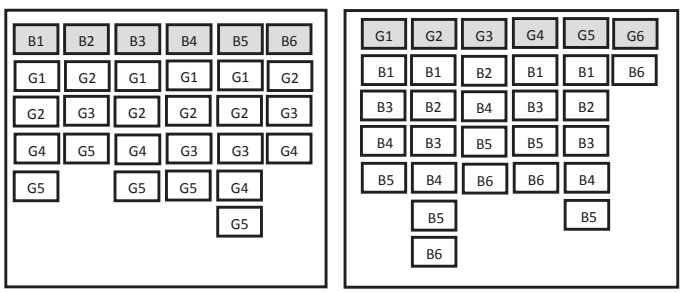
This task produces closed frequent item sets by implementing RISS algorithm Fig.4. RISS algorithm uses the Row Set data format and the minimal bottom-up search space.

Row SetData Format (RS)

RS is a new vertical format. Let table 1, be a discretized gene expression table of biological Samples B. {b1, b2 ,……..bum} in rows ,and genes G-{g1 ,g2 ,…….gn} in columns. Table 1 is a triple (B, G, R) ,where R B*G is a relation. (bi, gj) R denotes that gene gj is over expressed or under expressed in bi. Fig.2-a shows b1 has genes g1, g2, g4, and g5 are over expressed while g3 and g6 is under expressed.

Table 1: Gene expression table.

Biological Samples	g1	g 2	g 3	g4	g5	g 6
b1	1	1	0	1	1	0
b2	0	1	1	0	1	0
b3	1	1	0	1	1	0
b4	1	1	1	0	1	0
b5	1	1	1	1	1	0
b6	0	1	1	1	0	1



(a) RS (b)VTL

Figure 3. Comparison between the VTL and RS presentation of table 1.

The difference between VTL and RS is that the second uses the Tad (biological samples) instead of using the items (Genes) to build the VTL with condition that infrequent genes are deleted from the dataset. Fig.3 (a and b) shows RS and VTL of the data, in Fig.2 with mines up equals 30% of the samples.

The common genes among given samples is denoted as G (B'). Given a set of biological samples B'⊂ B, the row support set,denoted G (B›) ⊂ G is defined as the maximal set of genes common to all the rows in B› and the support of these genes denoted sup(G(B›)) is the count of Biological samples in B› As an example, let B›={b1, b3} in Fig.2 , then G (B›) = 2.

Minimal Bottom-up Search Strategy

It checks row combination from the smallest to the largest to traverse the search space, Fig.3-a, for example, 1- russets, then 2-rowsets, ..., and finally n-russets. It is a modification from the traditional one [5 ,6 ,7,8] as it starts the search from the mines up -russets rather than the 1- russets. This is valid as the maximum support for the K-russets ,for example if the mines up =3 the enumeration tree starts from level 3 as shown in Fig.3-b. Each node in the enumeration tree is represented by RS format. For example, the first node in Fig.3-b tree is represented by B'={1, 2, 3} and G(B') ={G2,G5-} as shown in Fig.3-c.

RISS Algorithm

RISS traverses the Minimal Bottom-up row enumeration search space using breadth first search (BPS). It has two pruning actions to eliminate unnecessary searches. RISS performs recursive generation of the Russets to present a node during traversing the row enumeration tree, Fig.4.

The first pruning action is performed in step 2 of the procedure intersection rows, it checks if the G(B') is empty. This implies that there is no maximal genes common to the biological samples B ,so no further enumeration will be required on the branch of this node.

The second pruning action is performed in step 3. It checks if the G (B') exists in CFI, if it is true, current and further enumeration of this node is truncated .In other words ,G(B') does not discover any new closed frequent item sets. At B'= {1 2 3}, then G (B') = {g2 g5} and results in {g2 g5) with frequency {1 2 3} e CFI. At B'= {1 2 4}, then G (B') = {g2 g5} which already exists in CFI and results in updating it to {g2 g5} with frequency {1 2 3 4}E CFI and no further enumeration for node {1 2 4}. At B'= {1 2 5), then G (B') = {g2 g5} which already exists in CFI and results in updating it to {g2 g5) with frequency {1 2 3 4 5}€ CFI and no further enumeration for node {1 2 5}. At B'= {2 3 4}, then G (B') = {g2 g5} which already exists in CFI and no further enumeration for node {2 3 4}. The same for B'. {2 3 5}. RISS does not need to perform the closure check among the discovered Item sets since step 1 in Intersection_ Rows procedure extracts only the closed item sets. The proof is that G (B') cannot be a maximal gene set that is common to all biological samples B' unless it is a closed item sets [10] . For example, in Fig. 2, it is not possible for the item set {g1 g2} to be enumerated although both {g2} and {g1 g2 g5} are closed item sets. This is unlike the column enumeration algorithms which enumerates both {g1 g2} and its superset {g1 g2 g5), then checks for equal support (= 4) and same frequency ({b1 b3 b4 b5)) between them. Thus, discarding the subset {g1 g2} and denoting the superset as a closed item set.

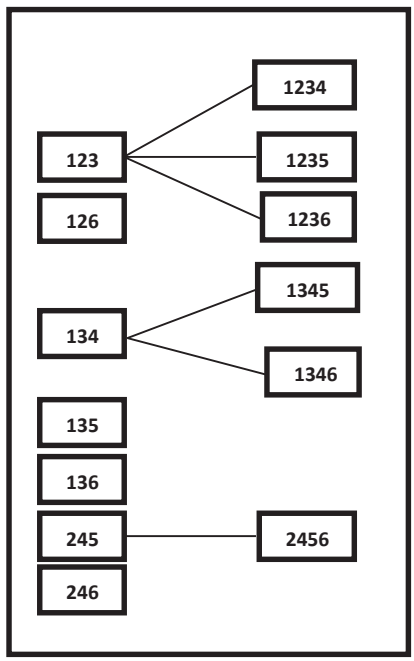


Figure 5.Pruning Enumeration Tree of Fig.3-b Tree.

b) Rule Discovery

Rule discoverytask is achieved by implementing the faster algorithm in [18] . It first generates all the rules with one item in the LHS. Then, recursively use the LHS with K items of the discovered rules to generate all

possible consequents with K+1 items that can appear in the further rules. All resulted rules with the associated parameters are stored in a database to ease querying and maintenance.

Experimental Study

Data

Array data is used of 24,483 gene measurements recorded for 19 breast cancer patients [19]. It contains: Systematic name given to each gene or sequence and a description of what is known about gene's function. Also, it contains three values for each tumor sample profiled: Log 10 (Intensity), Log10 (ratio), and P-value. According to the biology specialists' opinion, the P-value is used in mining process [20].

Experiments

Fifteen experiments are conducted with different filtration and minimum support thresholds. In the first step of filtration, removes the continues P-values keeping only the overall gene P-value. The second step uses three thresholds for the P-values which are 0.04, 0.05, and 0.06. Each filtered data is experimented by four mines up values (15.7%, 26.2%, 42.1%, 52.63%, and 63.15%) and mincing equals to 0 05. Experiments are performed on a PC with Core Duo 2 GHz CPU, 1GB RAM and a 120GB hard disc and the algorithm is coded in MATLAB.

Results

The results recorded for each experiment are shown in Table 1. For all experiments, the drawing charts exhibit decaying exponential function with different decaying parameter. Also, the drawn points collapse at higher support level for all datasets .In an attempt to carry the analysis of results one step further, a comparison for CFI and rules count is conducted. A General observation is that whenever the minus increases the common CFI and rule count among the sets increase except for one case AB in rule count at minus=8, the intersection reaches it.1 maximum then started to decrease with increasing mines up. Also, the percent of increase in intersection varies descending from AB to BC to AC.

Table 1: The experiments results

P value	Gene	Mines up	Processing Time (last three tasks)	CFI count	CFI Processing Time	Rules Count	Rules Processing Time	Total Processing Time
0.04	74	3	0.047	10663	1034.563	1804	2.125	1036.734
		5	0.047	7974	859.953	638	0.703	860.703
		8	0.063	2094	701.484	484	0.375	701.922
		10	0.063	624	624.609	448	0.297	624.969
		12	0.063	168	564.125	407	0.297	564.484
0.05	78	3	0.031	12303	1289.781	3796	2.313	1292.125
		5	0.047	9324	1045.422	801	0.656	1046.125
		8	0.031	2304	757.297	481	0.406	757.735
		10	0.047	656	655.172	434	0.313	655.531
		12	0.016	172	588.344	399	0.313	588.672
0.06	86	3	0.125	15197	1764.563	2158	2.906	1767.594
		5	0.078	12027	1468.609	1070	1.016	1469.703
		8	0.094	3301	1029.453	538	0.391	1029.938
		10	0.063	980	767.766	437	0.313	768.141
		12	0.063	253	616.594	524	0.313	616.969

Discussion

The proposed BCA system discovers the relationships among breast cancer patients' gene expressions. The first set of BCA advantages exists in the different preprocessing tasks. Data filtering is an important preprocessing task since it reduces the amount of data genes under study which speeds up the mining process, and gives more insight into specific data values. The first step in the data filtration removes data redundancy existing in the continues of the genes listed in the data file. Moreover, the second filtering step reduces the amount of the data which greatly enhances the performance in terms of processing time and memory space. The second task in the preprocessing, data adaptation is a necessary one in order to indicate the state of the gene in the rules. If not applied, the discovered 'rules specify only the over expressed associations. The discretization task converts the continuous data into binary one, thus speeds the processing cycle and reduces the memory requirement. This step did not affect the aim of the study since the interest is the differentiation between over and under expressed genes without using quantified measures.

Financial Disclosure: There is no financial disclosure.

Conflict of Interest: None to declare.

Ethical Clearance: All experimental protocols were approved under the University of Information Technology and Communications, Iraq and all experiments were carried out in accordance with approved guidelines.

References

1. Hancock, M Zvelebil, "Dictionary of Bioinformatics and Computational Biology. " A John Wiley & Sons, Inc., Publication, 2004.
2. Ritchie A. Bioinformatics Approaches for Detecting Gene—Gene and Gene—Environment Interactions in Studies of Human Disease. Neurosurg. Focus, October, 2005; 19.
3. Brazma J. Gene expression data analysis," European Molecular Biology Laboratory, Outstation HinXyon - the European Bioinformatics Institute, Cambridge CB 10 1SD, Uk.2000.
4. C Tang A. Mining Multiple Phenotype Structures Underlying Gene Expression Profiles," CIKM03, New Orleans, Louisiana, USA. 2003; 3-8.
5. A Sharaf, M Hana, T Soliman, S Rashad, "A Comparative Stud of Association Rules for mining Gene Expression Databases ,"International Journal of Intelligent Computing and Information Science, 2007; 7.
6. Ceglar J. Association Mining," ACM Computing Surveys, July, 2013; 38(2): Article 5,.
7. J Wang, J Han. Searching for the Best Strategies For Mining Frequent Closed Itemsets, "Proc.,2003 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining(KDD03), 2003.
8. M Zaki, C.Hsiao. Charm : An Efficient Algorithm for Closed Association Rule Mining," Proc. Of SIAM on Data Mining.2002.
9. J Pie, J Han, R Mao. CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets, Proc , 2000 ACM-SIGMOD Int. Workshop Data Mining and Knowledge Discovery(DMKD 00). 2000; 11-20.
10. F Pan, G Cong, A Tung, J Yang, M Zaki. Carpenter: Finding Closed Patterns in Long Biological Datasets," Proc. 2003 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'03).2003.
11. Pan, G Cong, X Xin, A Tung, "COBBLER: Combining Column and Row Enumeration for Closed Pattern Discovery,' Proc. the 16th International Conference on Scientific and Statistical Database Management (SSDBM'04), 2004 IEEE.
12. G Cong, A Tung, X Xu, F Pan, J Yang , Farmer: Finding interesting rule groups in microarray datasets," 23rd ACM International Conference on Management of Data, 2004.
13. G Cong, K Tan, A Tung, X Xu, "Mining Top-k Covering Rule Groups for Gene Expression," SIGMOD, 2005, Baltimore, Maryland, USA. 2005; 14-16.
14. P Shenoy, J Haristsa, S Sudatsham, G Bhalotia, M Baqa, D Shah, Turbo-charging Vertical Mining of Large Databases," Proceedings of the ACM SIGMOD (Austin, Texas), May, 2000; 22-29.
15. Rahal D Ren, A Perera, H Najadat, R Rahhal, W Valdivia, "Incremental Interactive Mining of Constrained Association Rules from Biological Annotation Data with Nominal Features," SAC'05, March 13-17, 2005, Santa Fe, New Mexico, USA.

Copyright 2005 ACM 1-58113-9640/05/0003.

16. M Zaki, K Gouda, "Fast Vertical Mining Diffsets," RPI Technical Report 01-1. Rensselaer Polytechnic Institute, Troy, NY 12180 USA. New York, 2001.
17. G Cong, K Tan, A Tung, F Pan, "Mining Frequent Closed Patterns in Microarray Data," Fourth IEEE International Conference on Data Mining (ICDM'04), 2004;363-366.