ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Frequency Distribution Mining Framework using OPSM

[I]J.Jeejo Vetharaj, [II]S.Jeevitha
[I]PG Scholar, Kalasalingam Institute of Technology, Krishnankovil, India
[II]AssistantProfessor, Kalasalingam Institute of Technology, Krishnankovil, India

Abstract
Discriminative patterns are patterns from data which have different frequency of distribution in each of the classes taken into comparison. These patterns reveal information from high dimensional gene data that can serve as a useful insight to differentiate the characteristics expressed due to genes in different animals or humans. This can be done by using support max pair techniques in a reasonable time span and with reasonably good accuracy. But in support max pair there are some drawbacks which add to the time of execution of the data in case of high dimensional data. This can be overcome effectively by using techniques OPSM to process the data and use supmaxpair to mine the discriminative patterns from the high dimensional gene data. OPSM technique groups similar data together allowing the frequent patterns to be discovered effectively. Using OPSM technique, data can be mined about 20 % lesser time compared to the existing method.

Keywords
Discrminative patterns, Apriori Algorithm, OPSM

I. Introduction

Discriminative patterns help in providing the difference between two or more classes in the form of patterns. They are the patterns that have a certain frequency of occurrence in one class and a completely disproportionate frequency in another class [1]. The explanation of discriminative patterns can be given with figure 1. In the figure 1 there are four patterns p1, p2, p3, p4 and two classes namely class 1 and class 2. Analyzing the patterns in the class 1 and class 2 of the dataset, a conclusion can be made that the patterns p4 is discriminative. Discriminative means the frequency of occurrence in the class 1 is higher than the frequency of occurrence in the class 2. But in case of pattern p2, the frequency of occurrence in the class 1 is very identical to the one in the class 2 so it is not classified as a discriminative pattern. Comparing this with the gene expression data, the pattern that leads to heart attack may be in class 1 and that does not cause may be in class 2 and the pattern can give an information that one class can signify the gene sequence or the pattern that is responsible for causing the heart attack. The difference however with the support between the classes is calculated that make the difference in both the ability to compute the patterns quicker and in an efficient manner. And this is very similar to the apriori algorithm and other association rule algorithms and it uses these properties in the computation too [1]. The difference is how ever with the support between the classes is calculated that make the difference in both the ability to compute / discover the patterns quicker and in an efficient manner. On using traditional algorithms on such data get inaccurate results in an unreasonable time span. And when dealing with data like gene data which are bigger in terms of both density and dimensionality the time factor in these cases must be given importance. It may take even days to complete finding patterns in the dataset. And again in case of gene data set it is important to get high accuracy so that wrong decisions which can prove costly in case of medicine or bioinformatics are not taken.

In gene expression analysis, the Order-Preserving SubMatrices (OPSMs) are employed to discover significant biological associations between genes and experiment conditions. The gene expression data are usually presented as a matrix in which the rows correspond to a set of genes, the columns correspond to a set of experiment conditions, and the entries represent the expression levels of the genes under the conditions [2].

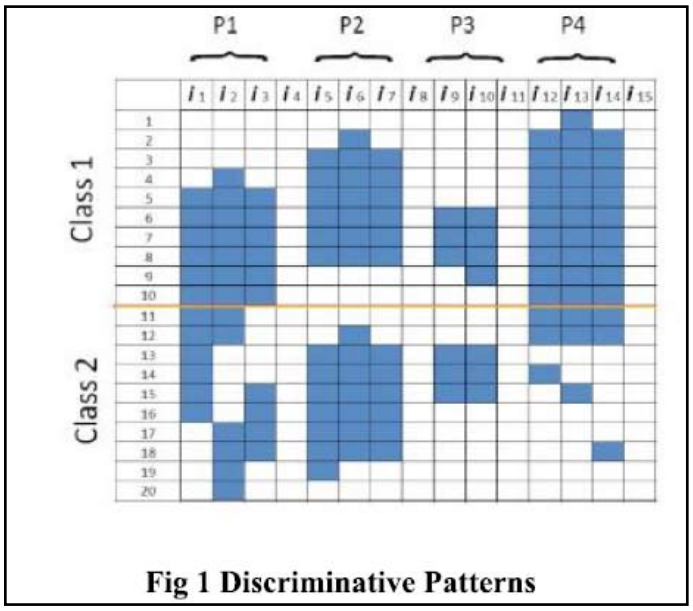

Fig 1 Discriminative Patterns

This model aims to capture the fact that the expression levels of a set of genes follow the same trend under a set of conditions.

II. Preprocessing of Data

The data used in this case is the gene expression data set where the discriminative pattern has been used extensively and this data is usually not used in the raw format. And for this a process called data discretization is carried out where a threshold value is taken into consideration. The Gene Expression data can be categorized in to two types they are the Expressed condition and the repressed condition. And in this case use of these states to transform the data to a format which is more similar to a transaction database and which is used in the Apriori and other pattern mining frameworks. The threshold is used to classify whether the data is expressed or repressed. The threshold is specific to the experiment and the gene under examination. If the data is over the threshold the data under consideration is considered to be the expressed under the experimental condition. And thus the data is taken to be 1 and else if it is not expressed it is taken to be 0. Thus transforming the data into a transactional database and this also serves as the input for the SMP framework which is used to mine low support discriminative patterns from data sets like gene expression data which have high dimensionality and high density.

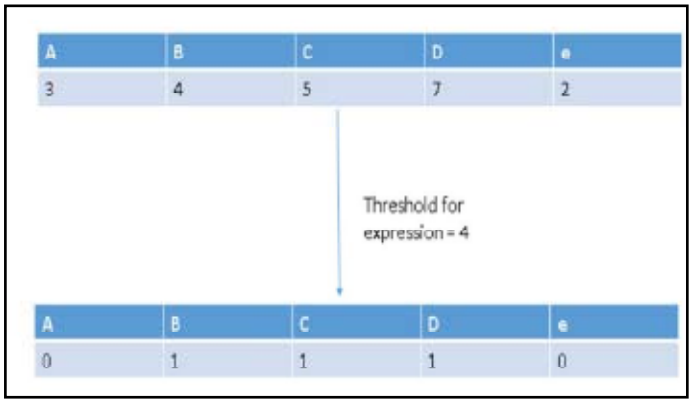

Fig. 2: Data Discretization

A. Candidate Generation

This is the first step of OPSM pattern mining and here the various patterns are generated .Here the data to be given as input to the SMP which works on the MAXSUPPORT of the size two patterns and thus a constrain to limit the size to a size two sub pattern is used. Here an algorithm which can generate top k patterns and this works on certain approximations to generate candidates. The accuracy can be improved with generating more patterns and validating the thus generated patterns. This algorithm takes k as input which is the no of patterns that must be generated.

Algorithm Generate candidate:
Step 1: Sort the support matrix.
Step 2: Repeat till K.
Step 3: keep the top columns of the sorted matrix to be fixed.
Step 4: Repeat till array length /2
Step 5: Swap randomly in the remaining array.
Step 6: Generate sub patterns of size 2 from the resulting array if it qualifies the inter and the intra bucket gap.

B. Candidate Validation

The generated patterns generated are validated against the rows of the data .Two thresholds are taken here one is the minimum number of rows or the percentage of rows and other is the minimum number of columns or the percentage of columns. That qualify the pattern. The pattern which qualify the most number of column is taken as the best pattern which is used further. The thus formed pattern is the OPSM. The algorithm given below is used to validate the pattern.

Algorithm

Step 1: For each generated candidate
Step 2: For each row in the data
Step 3: counter = 0; per=0;
Step 4: if (arr[counter]==1 && arr[counter+1]==1)
Step 5: per =per +1;
Step 6:if(per== min columns)
Step 7: return true
Step 8: Else return false.

Here the minimum columns is a user defined threshold. And the arr is an array with the key which is unique to each column of the pattern data set. The no of rows that qualify the patterns are stored as in a matrix. From the stored matrix the pattern with the maximum number of rows that qualify the column is taken as the best pattern. The input data is then rearranged to form a new data set with a new column order for the easier manipulation in the next process.

C. Discriminative Pattern Discovery

The first step is an algorithm-specific step. For example, for SupMaxPair, all the itempair supports are computed and stored in a matrix, whose entry is the item-pair support of items i and j. The complexity of this step is O (nm2), where n is the number of transactions, and m is the number of unique items. No such precomputation has to be done for CSET. The Apriori framework is said in this step for discriminative pattern mining using the antimonotonic measures BiggerSup and upMaxPair. For SMP, discriminative patterns are first mined from one class and then mined from the other, while CSET discovers patterns once from the whole data set. To facilitate further pattern processing and pattern evaluation, only the closed item sets from the complete set of item sets produced.

III. Apriori Algorithm With Supmaxpair

To find the discriminative patterns SupMaxK is used. The set of discriminative patterns organize themselves into nested layers of subsets. These nested layers are progressively complete in their coverage, but require more computation for their discovery. Given the same measure of time, the parts of this family furnish a tradeoff between the capacity to look for low-help discriminative examples specifically, an extraordinary part with K = 2 named Supmaxpair. This SMP helps for discovering discriminative patterns. Painstakingly designed experiments with both synthetic datasets and a cancer gene expression dataset are used to demonstrate that SMP can serve a complementary role to the existing approaches by discovering low-support yet highly discriminative patterns from dense and high-dimensional data, while the latter fail to discover them within an acceptable amount of time. Apriori is a classic algorithm for learning association rules [4]. Every set of data has a number of items and is called a transaction. As a result the output of Apriori is sets of rules that tell us how often items are contained in sets of data. The list of frequent itemsets generated during the first phase is scanned. If the list is empty, the procedure stops. Otherwise, let B be the next itemset to be considered, which is then removed from the list. 2. The set B of objects is subdivided into two non-empty disjoint subsets L and H = B − L, according to all possible combinations. 3. For each candidate rule $L \Rightarrow H$, the confidence is computed as p = conf $\{L \Rightarrow H\}$ = f (B) f (L). 4. If p ≥ pmin the rule is included into the list of strong rules, otherwise it is discarded. SupMaxK of an item set a is computed as the difference between the support of alpha in D1, and the maximal support among all the size-K subsets of alpha in D2. Note that, in this paper, Supmaxk is characterized concerning Diffsup, while comparable idea can likewise be connected to other discriminative measures, for example the degree based measure.
Apriori-gen function to determine the candidate itemsets before the pass begins. The interesting feature of this algorithm is that the database D is not used for counting support after the first pass. Rather, the set Ck is used for this purpose. Every member of the set Ck is of the form <TID; $\{X_k\}$ >, where each X_k is a potentially large k-itemset present in the transaction with identifier TID. For k = 1, C1 corresponds to the database D, although each item I is replaced by the itemset. For k > 1, Ck is generated by the algorithm. The member of C_k corresponding to the transaction t is <T ID, {c belongs to C_k} contained in it. If a transaction does not contain any candidate k-itemset, Ck will not have an entry for this transaction.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Thus, the amount of sections in C_k may be smaller than the amount of transactions in the database, particularly for extensive qualities of k. What's more, for expansive qualities of k, every passage may be more modest than the comparing transaction on the grounds that not many hopefuls may be held in the transaction. On the other hand, for little qualities for k, every section may be bigger than the relating transaction since an entrance in Ck incorporates all applicant k-itemsets held in the transaction.

IV. Results

This is a Graph drawn between different levels of smp and Diffsup with the corresponding support values for each of the pattern found. From this inferences can be made they are:

1. SM1 is a poor approximation of Diffsup since it has negative values and the input database is optimized to have no values less than 0.1.
2. SMP has a good approximation to the Diffsup and it does not have inaccurate values as the SM1.
3. To have a minimum computation effort on large and high dimensional data, SM2 is most suitable.

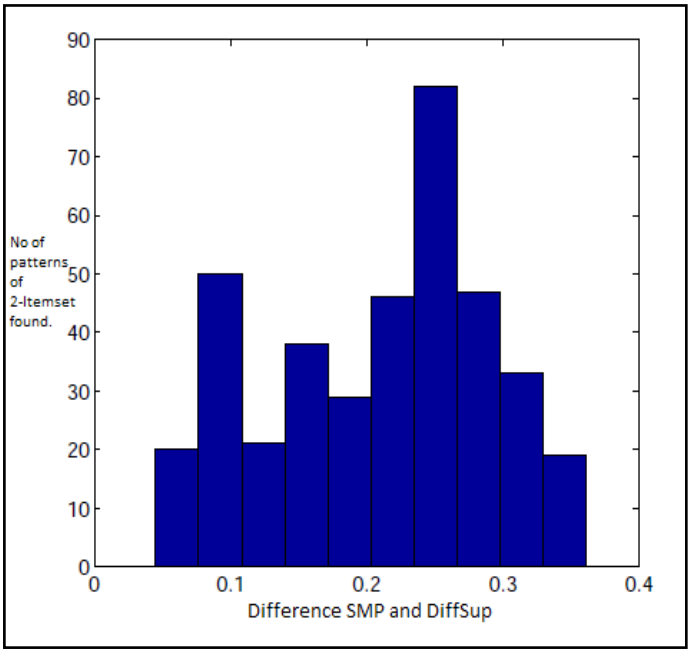

Fig. 3: Performance Graph

V. Conclusion and Related Work

This paper deals with the problem of the completeness of discriminative pattern discovery, with the ability to discover low-support discriminative patterns from dense and high-dimensional data within an acceptable amount of time. For this, a family of antimonotonic measures of discriminative power named SupMaxK that conceptually organize the set of discriminative patterns into nested layers of subsets, and are progressively more complete in their coverage, but require increasingly more computation for their discovery. Given the same and fixed amount of time, the SupMaxK family provides a trade-off between the ability to search for low support discriminative patterns and the coverage of the space of valid discriminative patterns for the corresponding threshold. It also deals with the problem of calculating supports in the framework to overcome this a technique called OPSM which arranges the data in form of patterns which have uniform distribution throughout the dataset and in turn we give this uniformly distributed data as input to the algorithm as a result we are able to discover the patterns in a reasonable amount of time. The improvement in the time of Execution of the algorithm is found to be about 20 % from the model where the unprocessed data is given as an input. And in future the algorithm can be further improved by the use of efficient data structures for pattern mining in the framework to find the discriminative patterns.

References

[1] *Gang Fang, GauravPandey, Wen Wang, Manish Gupta, Michael Steinbach, Member and Vipin Kumar IEEE transactions on knowledge and data engineering, vol. 24, no. 2, February 2012.*

[2] *D. Segre et al., "Modular Epistasis in Yeast Metabolism," Nature Genetics", vol. 37, pp. 77-83, 2004.*

[3] *C. Carlson et al., "Mapping Complex Disease Loci in Whole genome Association Studies," Nature, vol. 429, no. 6990, pp. 446-452, 2004.*

[4] *S. Bay and M. Pazzani, "Detecting Group Differences: Mining Contrast Sets," Data Mining and Knowledge Discovery, vol. 5, no. 3, pp. 213-246, 2001.*

[5] *P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Addison-Wesley, 2005.*

[6] *RakeshAgrawalRamakrishnanSrikant Fast Algorithms for Mining Association Rules IBM Almaden Research Center.*

[7] *R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proc. Very Large Data Bases (VLDB), pp. 487-499, 1994.*

[8] *N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering Frequent Closed Itemsets for Association Rules," Proc. Int'l Conf. Database Theory (ICDT), pp. 398-416, 1999.*

[9] *A. Soulet et al., "Condensed Representation of Emerging Patterns," Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD), pp. 127-132, 2004.*

[10] *A. Subramanian et al., "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles," Proc. Nat'l Academy of Sciences USA, vol. 102, no. 43, pp. 15545-15550, 2005.*

Jeejo Vetharaj J is currently pursuing M.E in Kalasalingam Institute of Technology under Anna University Chennai. His research interest includes Data mining and Weka tool.

Jeevitha S is currently working as Assistant professor in Kalasalingam Institute of Technology, Krishnankovil. She has an experience of four years in Teaching. Her Research area includes Data Mining and Image Processing.