Applied Soft Computing 38 (2016) 587–605

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Fuzzy Correlated Association Mining: Selecting altered associations among the genes, and some possible marker genes mediating certain cancers

Anupam Ghosh [a], Rajat K. De [b,*]

[a] Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata 700152, India
[b] Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India

ARTICLE INFO

Article history:
Received 23 December 2014
Received in revised form 8 August 2015
Accepted 27 September 2015
Available online 17 October 2015

Keywords:
Transcriptional regulation
p-Value
Biochemical pathways
Functional enrichment

ABSTRACT

Association mining is a well explored topic applied to various fields. In this article, the associations among the genes have been identified from microarray gene expression data. Here a methodology, called Fuzzy Correlated Association Mining (FCAM), is developed for identifying the associations among the genes that have altered quite significantly from normal state to diseased state with respect to their expression patterns. This idea leads to predict the disease mediating genes along with their altered associations. The proposed methodology involves generation of fuzzy gene sets, construction of fuzzy items, computation of fuzzy support for fuzzy items and fuzzy correlation coefficient of a pair of fuzzy items, generation of associations, and identification of altered associations from normal to diseased state. The concept of finding fuzzy correlation between two groups of items, generation of altered associations among the items (groups of items) and then rank these items (groups of items) according to their importance are the novel contribution of the present article. The effectiveness of the methodology has been demonstrated on five gene expression data sets dealing with human lung cancer, colon cancer, sarcoma, breast cancer and leukemia. As a result, some possible genes, like IGFBP3, ERBB2, TP53, HBB, KRAS, PTEN, CALCA, CDKN2A, has been found as important genes that may mediate the development of various cancers considered here. For comparison, we have considered 11 existing association rule mining algorithms. The results are appropriately validated in terms of gene–gene interactions, functional enrichment, biochemical pathways, and using NCBI database.

1. Introduction

Exploratory data mining techniques are needed that can, roughly speaking, be considered as the search for interesting bi-sets, i.e., sets of biological situations and sets of genes, which are associated in some way. Indeed, it is interesting to look for groups of co-regulated genes, for which a reasonable assumption is that they participate in a common function within the cell [1]. Genes are grouped together according to similar expression profiles. The association among a set of co-regulated genes and its discovery pave a way to a better understanding of gene regulation.

Fuzzy set theory is capable of handling uncertainty in the gene expression values arising due to incompleteness, imprecision, noise and experimental errors. Moreover, genes have expression values that are in different intervals under two conditions (i.e., normal or diseased). Although each interval has a well-defined boundary, they are highly overlapped. Fuzzy set theory is especially suitable to model such imprecise and overlapping data. Thus incorporation of the notion of fuzzy sets in the methods enables one to handle such overlapping intervals in a better way [2,3].

The notion of fuzzy sets has been used in the domain of gene expression analysis. They include, among others, development of rule discovery procedure [4] based on knowledge extraction of gene by classification; transformation of gene expression by fuzzy heuristic rule set [5]; classifying fuzzy inference system [6]; development of a fuzzy model for gene regulatory networks [7]; measuring performance of small rule-based classifiers using fuzzy logic [8]; identification of normal and tumor patients using a fuzzy neural network model [9].

Global gene expression profiling, both at the transcript level and at the protein level, can be a valuable tool in the understanding of genes, biological networks and cellular states. As larger and larger gene expression data sets become available, data mining techniques can be applied to identify patterns of interest in the data. Association rules, used widely in the area of market basket analysis, can be applied to the analysis of expression data as well. Association rules can reveal biologically relevant associations between different genes or between environmental effects and gene expression [10]. Items in gene expression data can include genes that are highly expressed or repressed, as well as relevant facts describing the cellular environment of the genes (e.g., the diagnosis of a tumor sample from which a profile was obtained).

Association rule discovery has been applied to gene expression data, searching for patterns of differential expression across tens of thousands of genes. In

* Corresponding author. Tel.: +91 3325753105; fax: +91 3325783357.
E-mail addresses: anupam.ghosh@rediffmail.com (A. Ghosh), rajat@isical.ac.in (R.K. De).

http://dx.doi.org/10.1016/j.asoc.2015.09.057
1568-4946/

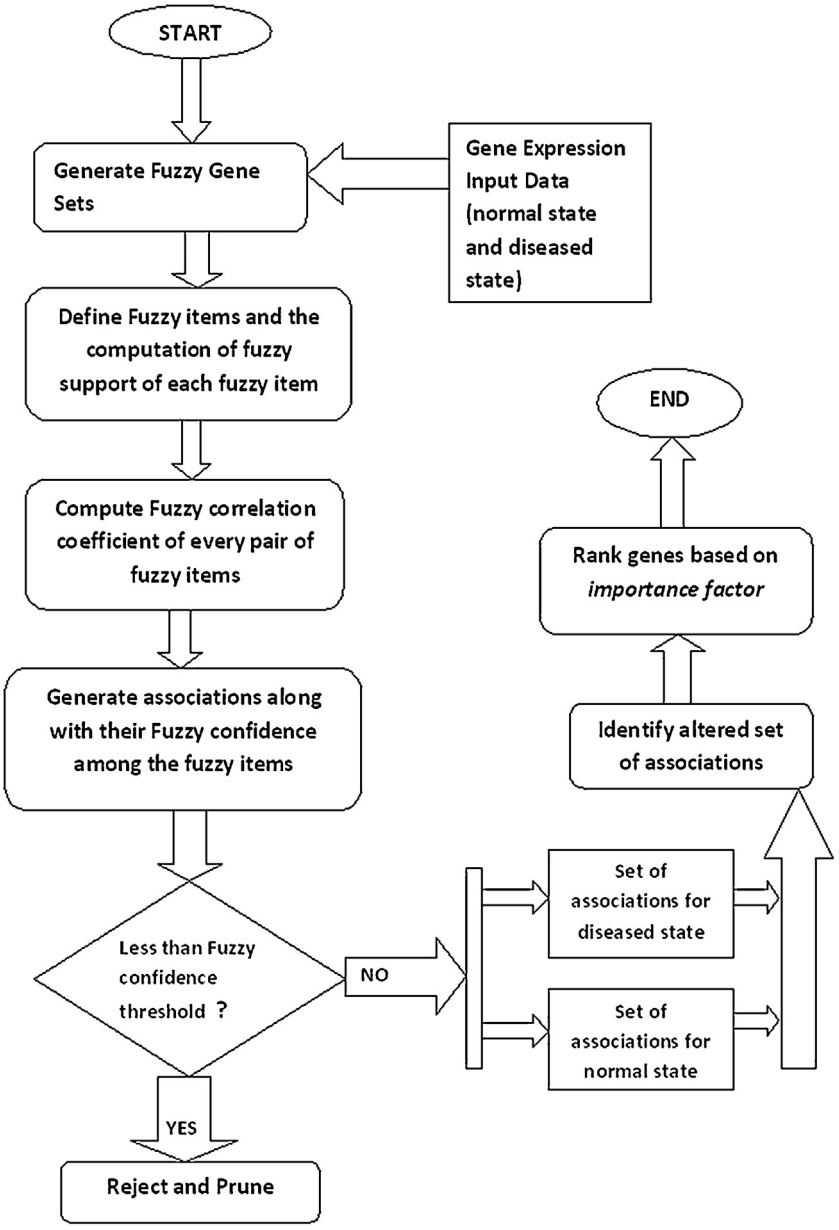

Fig. 1. Flowchart of FCAM.

life threatening diseases, such as cancer, where the effective diagnosis includes annotation, early detection, distinction, and prediction, data mining and statistical approaches offer the promise for precise, accurate, and functionally robust analysis of gene expression data [11]. The computational extraction of derived patterns from microarray gene expression is a non-trivial task that involves sophisticated algorithm design and analysis for specific domain discovery. In an earlier investigation, a model was proposed for feature extraction by first applying feature selection heuristics based on the statistical impurity measures like Gini Index, Max Minority, and the Twoing Rule and obtaining the top 100–400 genes and then analyze the associative dependencies between the genes and assign weights to the genes based on their degree of participation in the rules [12]. An analysis of some of these rules reveals numerous associations among certain genes, many of which make sense biologically, others suggesting new hypotheses that may warrant further investigations.

In this article, an association rule-mining algorithm, called fuzzy correlated association mining (FCAM), has been developed, which uses gene expression data to generate the association among the large set of genes and more importantly to identify the set of altered associations. The algorithm considers a large set of genes and determines the disease mediating genes (marker genes) (Fig. 1).

Unlike the proposed algorithm FCAM, there is no method that determines gene–gene associations that have altered from normal state to diseased one and thereby finding possible disease mediating genes. Thus, identifying the associations among the genes, discovering the altered associations from gene expression data and finally the proposal of gene ranking technique that is used to identify the importance of the genes from the set of altered associations may be considered as a novel concept. Biological data are often imprecise and noisy. Moreover, in microarray gene expression data, genes have expression values that are in different intervals under two conditions, i.e., normal or diseased. Although each interval has a well-defined boundary, they are highly overlapped. Fuzzy set theory is especially suitable to model such imprecise and overlapping data. This idea leads us to develop the methodology using the concept of fuzzy set theory. Thus incorporation of the notion of fuzzy sets in the proposed method enables one to handle such overlapping intervals in a better way.

We have applied the proposed methodology (FCAM) (Fig. 1) on gene expression data sets of 5 different human cancers (lung, colon, lymphocyte, sarcoma and breast). There exist several methods for finding gene–gene associations. These existing methods may be classified into 3 categories, *viz.*, Category 1, Category 2 and Category 3. The methods under Category 1 consider finding gene–gene associations from structural data, i.e., sequence data [13–15]. Under Category 2, the methods determine associations among gene functions and protein–protein interactions by using gene expression data as well as some other databases [16–20]. The methods, under Category 3, determine association among objects in a non-biological domain [21–26]. Since we could not find any method that works exactly the same way as FCAM does, we have considered 11 methods under Categories 2 and 3 for comparison. Although it can extract the associations among the