ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 13, No. 3, 2018, *pp*.168-178

Fuzzy Discretization and Rough Set based Feature Selection for High-Dimensional Classification

Prema Ramasamy [1], Premalatha Kandhasamy [2]

[1] *Prema Ramasamy, Assistant Professor, New Horizon College of Engineering, Bangalore*
E-mail:premabit@gmail.com
[2] *Professor, Department of Computer Science and Engineering, Bannari Amman Institute of Techlology, Sathyamangalam.*

(Received May 11 2018, accepted July 16 2018)

Abstract. Contemporary biological technologies like gene expression microarrays produce extremely high-dimensional datasets with limited samples. Analysis of gene expression data is essential in microarray gene expression studies in order to retrieve the required information. Gene expression data generally contain a large number of genes but a small number of samples. The complicated relations among the different genes make analysis more difficult, and removing irrelevant genes improves the quality of results. In this regard, a new feature selection algorithm called 2-level MRMS is presented based on rough set theory. It selects a set of genes from microarray data by maximizing the relevance and significance of the selected genes. The paper also presents a novel discretization method, Gaussian Fuzzy Discretization based on fuzzy logic to discretize the continuous gene expression values. The performance of the proposed algorithm, along with a comparison with other related feature selection methods, is studied using the classification accuracy of k-Nearest Neighbor (kNN) and Support Vector Machine (SVM) on four microarray data sets. The experimental results show that the genes selected using 2-level MRMS feature selection give high classification accuracy than other methods.

Keywords: classification, feature selection, fuzzy discretization, high- dimensional data, maximum relevance and maximum significance, microarray data

1. Introduction

A microarray dataset [1] is a repository containing microarray gene expression data. The raw microarray data are images that are transformed into gene expression data matrices where rows represent genes, columns represent various samples such as tissues or experimental conditions and the numbers in each cell characterize the expression level of a particular gene in a particular sample. Figure 1 shows an example of an $M \times N$ gene expression dataset where M is the number of genes, and N is the number of samples.

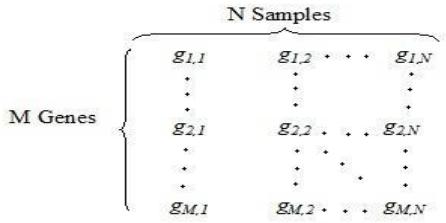

Fig. 1: Example of gene expression data

Data dimensionality reduction is one of the important machine learning tasks while dealing high-dimensional data with enormity on size, missing values and noise [2]. Gene expression dataset contains thousands of gene expression values, many of which may be irrelevant or redundant for classification [3]. Leaving out relevant attributes or keeping irrelevant attributes may affect the performance of the classification algorithm. Therefore statistical methods are required to identify a reduced search space are commonly used for classification [4]. There are many feature selection approaches to assist in classification of samples [5-9]. They are classified into four categories, namely filter approach, wrapper approach, embedded approach, and hybrid approach. A filter approach applies a statistical measure to assign a score to each feature without using

a learning algorithm [10]. A wrapper approach uses learning techniques to evaluate the accuracy produced by the use of the selected features in the classification [11]. An embedded approach combines the feature selection step and classifier construction. A hybrid approach is a combination of both filter and wrapper-based methods [12]. In this paper, Rough Set Theory (RST) based feature selection method is used.

The rough set theory has been applied successfully to feature selection of discrete valued data [13,14,15]. It reduces the number of features of a dataset without considering any prior knowledge and using only the information contained within the dataset [16]. In this paper, 2-level MRMS feature selection method is proposed to select a set of genes from gene expression datasets by considering both relevance and significance of the selected genes. To compute the relevance and significance, the equivalence partitions of each gene is used. This can be automatically derived from the given datasets. So, RST needs no information other than the data set itself.

The RST feature selection process can only operate effectively with datasets containing discrete values. As gene expression datasets contain continuous value attributes, it is necessary to perform a discretization technique before gene selection. This paper presents a new discretization method, Gaussian Fuzzy Discretization (GFD) to discretize the continuous gene expression values. The discretization of numerical attributes can be performed before or after normalization [l7]. In this paer, the datasets are normalized using fuzzy logic. Then the normalized dataset can be discretized using mean and standard deviation.

The GF-discretized datasets are given as input to the feature selection methods. The rest of the paper is organized as follows. Section 2 discusses the related work in brief. Section 3 introduces the basic concepts of rough sets. The GFD process and the proposed feature selection method is explained in Section 4 for selecting relevant and significant genes.

2. Related Work

Hu et al. [18] proposed a feature subset selection technique based on a fuzzy-rough model. They used a symmetric function to compute fuzzy similarity relations between the objects with a numerical attribute and transform the similarity relation into a fuzzy equivalence one. So, this approach does not require discretizing the numerical data. Also they defined four attribute significance measures. Based on the measures, they constructed a forward hybrid attribute selection algorithm. Jenson and Shen [19] examined a novel approach based on fuzzy-rough sets, called fuzzy-rough feature selection. It overcomes the problems of noisy and real-valued data, as well as handling mixtures of nominal and continuous value attributes. FRFS achieves this by the use of fuzzy-rough sets, and a new measure of attribute significance, the fuzzy-rough degree of dependency. It also deals with real-valued decision attributes.

Yao and Zhao [20] discussed attribute reduction in decision-theoretic rough set models regarding different classification properties, such as: decision monotocity, confidence, coverage, generality and cost. Cornelis et al. [21] introduced a framework for fuzzy-rough set based feature selection. They provided a comprehensive typology of subset evaluation measures that can be used to define fuzzy decision reducts. Zhang et al. [22] studied the attribute reduction based on a discernibility matrix and used it to design correspondence attribute reduction algorithm. A simplified decision table was first introduced and then, a new measure of the significance of an attribute was defined for reducing the search space of the simplified decision table. Zahra and Reza [23] proposed a new fuzzy 2-level complementary system for classification of gene expression data. This approach exploits complementary learning and hierarchical organization, and complexity reduction and good interpretability are achieved.

3. Rough Sets

Let $I = (U, A \cup D)$ be an information system [24], where U is a non-empty set of finite objects (the universe) and A is a non-empty finite set of attributes and D is the set of decision attributes. This information system can be called as decision table. $\forall a \in A$, there exists a corresponding function $f_a: U \to V_a$, where V_a is the set of values that attribute a take. If $P \subseteq A$, there is an associated equivalence relation [24]:

$$IND(P) = \{(x, y) \in U^2 \mid \forall a \in P, f_a(x) = f_a(y)\} \quad (1)$$

The partition of U, generated by $IND(P)$ is denoted U/P. If $(x, y) \in IND(P)$ then x and y are indiscernible to P. The equivalence classes of the P-indiscernibility relation are denoted as $[x]_P$. Let $X \subseteq U$, the P-lower approximation $\underline{P}X$ and P-upper approximation $\overline{P}X$ of set X can be defined as [24]:

$$\underline{P}X = \{x \in U | [x]_P \subseteq X\} \quad (2)$$