Applied Soft Computing 65 (2018) 400–411

Contents lists available at ScienceDirect

Applied Soft Computing

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

A group incremental feature selection for classification using rough set theory based genetic algorithm

Asit K. Das [a], Shampa Sengupta [b], Siddhartha Bhattacharyya [c,*]

[a] Department of Computer Science and Technology, Indian Institute of Engineering Science and Technology, Shibpur, Howrah 711 103, West Bengal, India
[b] Department of Information Technology, MCKV Institute of Engineering, Liluah, Howrah 711 204, West Bengal, India
[c] Department of Computer Application, RCC Institute of Information Technology, Kolkata, India

ARTICLE INFO

Article history:
Received 12 May 2017
Received in revised form 27 January 2018
Accepted 28 January 2018
Available online 3 February 2018

Keywords:
Data mining
Rough set theory
Genetic algorithm
Incremental data
Feature selection
Classification

ABSTRACT

Data Mining is one of the most challenging tasks in a dynamic environment due to rapid growth of data with respect to time. Dimension reduction, the key process of relevant feature selection, is applied prior to extracting interesting patterns or information from large repositories of data. In a dynamic environment, newly generated group of data together with the information extracted from the previous data are analyzed to select the most relevant and important features of the entire data set. As a result, efficiency and acceptability of the incremental feature selection model increase in the field of data mining. In our paper, a group incremental feature selection algorithm is proposed using rough set theory based genetic algorithm for selecting the optimized and relevant feature subset, called reduct. The objective function of the genetic algorithm used for incremental feature selection is defined using the previously generated reduct and positive region of the target set, concepts of rough set theory. The method may be applied in a regular basis in the dynamic environment after small to moderate volume of data being added into the system and thus the computational time, the major issue of the genetic algorithm does not affect the proposed method. Experimental results on benchmark datasets demonstrate that the proposed method provides satisfactory results in terms of number of selected features, computation time and classification accuracies of various classifiers.

1. Introduction

Recently, data generated specially in the field of Engineering Science are not robust; it comes in complex format with higher dimensions. So mining the important information from such data is a challenging job. The data size also increases with respect to time which motivates to develop a novel data analysis technology. Learning techniques concerned with the incremental approach [1–7] are frequently applied on dynamic data to discover knowledge in an effective and efficient way. A learning algorithm [8] is considered as an incremental learning algorithm if it generates a series of hypotheses $H_0, H_1, \ldots, H_n$ based on the training sets $T_1, T_2, \ldots, T_n$ respectively, where H_{i+1} depends only on H_i and the current training set T_i. However, the resultant hypothesis is applicable for all the training datasets available so far. So it reduces both the space and the time complexity with respect to storing and processing of data. An incremental algorithm [8] can detect the change in environment and adjust its hypothesis with the new environment. In modern days, feature selection, a very frequently used technique in data mining [9], has attracted much attention for efficient storing and retrieving of high dimensional data.

Feature selection [10] and reduct generation [11,12] are the preprocessing techniques in data mining [11–13] for discovering knowledge from the stored data. According to a certain assessment criterion, optimal subsets of features are selected from the entire feature set. It is an interesting field of research that has been exposed to be very efficient for removing irrelevant and redundant features. The importance of a feature subset is judged by parameters such as its *relevancy* and *redundancy*. A relevant feature is always predictive of the decision feature, otherwise features are considered as irrelevant. A feature is redundant if it is highly correlated with some other features. The main objective of our work is to find out an optimal subset of features that are uncorrelated with each other and extremely correlated with decision feature. Feature selection algorithms [14] can be broadly divided into three approaches such as filter approach, wrapper approach and embed-

* Corresponding author.
E-mail addresses: akdas@cs.iiests.ac.in (A.K. Das), shampa.sengupta@mckvie.edu.in (S. Sengupta), dr.siddhartha.bhattacharyya@ieee.org (S. Bhattacharyya).

https://doi.org/10.1016/j.asoc.2018.01.040
1568-4946/

ded approach based on their working principle. If an algorithm performs feature selection independently of any learning algorithm then it is a *filter* approach [14] where irrelevant features are filtered out before applying any classification algorithm [13]. On the other hand, if the feature selection process is tied with a learning algorithm, then it is classified as a *wrapper* approach and the suitability of the feature subset is evaluated by the estimated accuracy from a learning algorithm. Wrapper methods [14] can be classified as Sequential Selection Algorithms (SSA) and Heuristic Search Algorithms (HSA). The SSA works in a way of adding features (or removing features) in an empty set (or full set) until the maximum value of the objective function is reached. To select the features, a criterion is set so that maximum value of the objective function is achieved with the minimum number of features. HSA works in a manner where different feature subsets are evaluated in a search space to optimize the objective function. The feature subset providing maximum value for the objective function is selected as the optimal feature subset. Embedded methods [14] work in a manner where the feature selection part is integrated in the training process to reduce the total time taken up for reclassifying different subsets unlike wrapper approach. Our method is based on filter approach where a heuristic search algorithm such as Genetic Algorithm (GA) [15] has been used to find the optimal feature subset with a novel and suitable rough set theory based objective function.

Principal Component Analysis (PCA) [16], which transforms correlated variables into the principal components, is one of the popular methods for dimension reduction. PCA is extensively applied in computer vision [17] as it can extract the inclusive structure of a data set. Linear Discriminant Analysis (LDA) [18], a recognized dimension reduction algorithm, transforms the original data to a lower dimensional space by maximizing the ratio of interclass distance to intra-class distance. It is popularly used in text retrieval [19], face recognition [20], and microarray data classification [21]. PCA and LDA are generally used for static data, because the involvement of the Eigen value problem is difficult to maintain incrementally. Some research works [22,23] exist on vision and numerical linear algebra handled PCA for incremental data.

In most of the real life applications, datasets change with time that make the static data mining algorithms extremely inefficient for knowledge discovery from database as the same learning algorithm need to be run repeatedly. There are a number of rough set based research papers [11,12] that discuss different approaches of generating the reduct for static data or time invariant data. The attribute reduction method using standard rough set theory [24–26] is effective to some extent, but there are some problems that must be solved in practice, especially for the incremental data set, which is time variant in nature. The feature selection problem for incremental data basically lies in the category of online algorithms and hence a dynamic solution is needed to avoid the reuse of old data.

To handle the dynamic data, several incremental feature selection algorithms [27–30] had been proposed by the researchers. A common characteristic of those algorithms is that, they are appropriate for new data that is being generated one by one. When many objects are produced at a time, these algorithms may not be efficient enough, as repetitive execution is needed to handle the new group of objects. Guan [31] developed an incremental updating algorithm to find a reduced attribute set of the decision system based on the discernibility matrix [11,12], where new group of objects added in the decision system changes the discernibility matrix and accordingly the selected attribute set is modified. Hu et al. [32] developed an incremental attribute reduction algorithm based on the elementary sets, which can determine the reduced attribute set from a dynamic information system. Wang et al. [33] developed an attribute reduction algorithm for datasets with dynamic data values using the concept of information entropy. Deng [34] presented an attribute reduction method by generating a set of reducts in parallel using the concept of positive region of rough set theory and the significance of the attributes. Bazan et al. [35] introduced the concept of dynamic reducts to handle incremental data set, in which the quality of the dynamic reduct is measured using the stability coefficients. Jun et al. [36] developed an improved incremental attribute reduction algorithm by exploring the concept of relative positive region, which can handle both the incremental attributes and incremental samples. Liang et al. [37] proposed a group incremental method for feature selection in the framework of rough set theory. The method used information entropy as a parameter for measuring significance of the features.

Dun et al. [38] proposed a matrix based incremental approach in dynamic incomplete information system for knowledge discovery. In the method, three types of matrices namely support matrix, accuracy matrix and coverage matrix under four different extended relations such as tolerance relation, similarity relation, limited tolerance relation and characteristics relation were introduced to incomplete information systems for inducing knowledge dynamically. Though the method is helpful to deal with the missing and incomplete data, but it is time consuming for learning knowledge from datasets with high volumes, as addition and deletion of individual objects take place for knowledge discovery in this type of incremental model. Dey et al. [39] proposed an important method for discretization and reduct generation simultaneously. The method tried to find out more effective cut in each step and found out more cuts to select the attributes with higher significance. Cuts from more significant attributes are more effective and the number of cuts needed to consistently classify a sample of objects is proportional to the square root of the sample size. The method was applicable for static datasets but there were no specific discussion about the incremental datasets in the work. Moreover, the experimental results given were only for three datasets with very few numbers of attributes and decision classes. Xu et al. [40] proposed an incremental attribute reduction method based on 0–1 integer programming when multiple objects enter into an information system incrementally. The method updated the old reduct based on the newly entered data subset into the information system. Though the method is helpful for attribute reduction in incremental environment but the performance of the method is not very significant with respect to other incremental algorithms. Shu et al. [41] proposed a method for incremental feature selection which is very important for dynamic incomplete data. Method employed a rough set theory based incremental approach to compute the new positive region when objects with varied feature values were added dynamically. Based on the calculated positive region value, features were selected incrementally. The paper proposed two efficient incremental feature selection algorithms, one considering single new object at a time and the other considering multiple objects or group of new objects entering into the system. The paper explained that single new object based incremental feature selection technique is more time consuming and provides poor feature selection performance compare to multiple objects based incremental feature selection algorithm. So, our proposed method is compared with only the second approach of this paper.

In our work, an algorithm based on rough set theory [24] and genetic algorithm [15] is proposed for group incremental feature selection. The algorithm handles any incremental data for finding an optimized feature subset, called reduct by modifying the previous reduct whenever new group of data are added. The proposed algorithm has been applied on various benchmark datasets to demonstrate its effectiveness.

The paper is structured as follows: preliminaries on Genetic algorithm and rough set theory are discussed in Section 2. Section 3 demonstrates the proposed group incremental feature selection technique. Section 4 shows the experimental results and compar-