Neurocomputing 116 (2013) 367–381

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Reducing bioinformatics data dimension with ABC-kNN

Thananan Prasartvit [a], Anan Banharnsakun [a], Boonserm Kaewkamnerdpong [b], Tiranee Achalakul [a,*]

[a] Department of Computer Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi, Thailand
[b] Biological Engineering Program, Faculty of Engineering, King Mongkut's University of Technology Thonburi, Thailand

ARTICLE INFO

Available online 22 October 2012

Keywords:
Dimension reduction method
Artificial Bee Colony
k-Nearest Neighbor
Classification problems

ABSTRACT

Analyzing a large amount of data often consumes extensive computational resources and execution time. However, sometime all data features do not equally contribute to the end results. Thus, it is plausible to identify the major contributing features and use them as representatives of the data. Other features with low contribution can be eliminated to reduce the time/resource consumption in data analysis. One of the promising application domains for such a feature selection process is *bioinformatics*. The need for dimension reduction, which is the process to reduce unnecessary features from the original data, arises because biological data can be massive, with tens of thousands of features to be explored. The objective of this study is to design an effective algorithm that can selectively remove irrelevant dimensions from data describing complex biological processes while preserving the semantics of the original data. This research proposes the adoption of the Artificial Bee Colony (ABC) as a novel method for data dimension reduction in classification problems. ABC, an efficient heuristic method based on swarm intelligence, is used to select the optimal subset of dimensions from the original high-dimensional data while retaining a subset that satisfies the defined objective. The k-Nearest Neighbor (kNN) method is then used for fitness evaluation within the ABC framework. In this research, ABC and kNN have been modified and bundled together to create an effective dimension reduction method. The proposed algorithm is validated in two distinct application domains: Gene expression analysis, and autistic behaviors study. The experimental results exhibit good solution quality as well as good computational performance.

1. Introduction

The advance of computer technology has led us to the information era, where more and more data are available for analysis. Research in the field of bioinformatics has led to new understanding of complex biological processes that have direct impacts on our everyday life. Nevertheless, the data sets produced by biological research can be massive. In some studies [1–3], there can be tens of thousands of potentially interesting factors to consider. Such a vast amount of data could lead to large consumption of resources and execution time in the data analysis process.

Many bioinformatics studies can be considered as example of data mining, that is, discovering new knowledge or learning data pattern from large dataset. In data mining, it is usually the case that not all collected data contribute to the significant findings. Some of the data dimensions could possibly be excluded to reduce the computation complexity and consumption while the remaining data could maintain the vital information. Hence, dimension reduction is an important process in data mining for analyzing massively high-dimensional data.

One dimension reduction process is called *feature selection*. This process attempts to maintain the characteristics of the original data by removing non-essential data such as noise or redundancy. A feasible subset of features from the original set of candidate features is selected. For data mining problems, feature selection may be performed as the intermediate step in the evaluation process for the large dimensional data. It is used to maintain evaluation accuracy and increase efficiency of data mining methods.

Several techniques have been used in feature selection, for example, Branch and Bound (BB) algorithm. BB was first introduced in [4]. In the BB algorithm, a feature set is represented as a tree structure; Each node represents a feature. The BB method searches and evaluates all possible feature subsets based on a defined criterion. The BB method can efficiently discover several suboptimal subsets and discard them so that less evaluation time will be required. Nevertheless, BB performs a complete search for the optimum subset of features through all possible feature sets. Although the best subset of features that can satisfy the

* Corresponding author.
E-mail addresses: pra.thananan@gmail.com (T. Prasartvit), anan_cpe@yahoo.com (A. Banharnsakun), boonserm.kae@kmutt.ac.th (B. Kaewkamnerdpong), tiranee@cpe.kmutt.ac.th (T. Achalakul).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.01.045

predetermined criteria can be obtained, enormous computation consumption is inevitable.

Another feature selection technique is the stochastic method. This method randomly generates a subset of features that satisfy some defined parameters and then randomly adds optimal features (forward search) or removes redundant features (backward search) from the selected subset of features. In order to achieve the optimal result, evaluation functions are used to estimate the error of the subset of features. Compared to complete search techniques like BB, this technique can generally reduce execution times. However, when performed on massively large data, the stochastic method can lead to expensive computation. Moreover, it is not guaranteed that the optimal subset of features can be found.

Heuristic techniques are often applied in feature selection. These techniques explore and update a subset of features based on predetermined rules to find an optimal subset of features. Thus, these techniques require less computation than the previous techniques even when used in problems with massive amounts of data. Some heuristic techniques in the swarm intelligence algorithm family can even produce outstanding results in feature selection [5–7].

Even for problems with lower dimensionality of data, dimension reduction can be a costly process because the process requires many steps as well as computational time to compute and search for the optimal subset. For bioinformatics problems where biological data contain vital information about living organisms in high dimensional form, heuristic techniques seem more efficient and suitable to apply for effective feature selection. In bioinformatics, where the biological data are used to reveal the underlying process and to better understand biological systems, a feature selection method integrated with a classification method is usually employed. The accuracy of the selected subset of features can be evaluated with a supervised-learning classification method in order to learn and recognize patterns of data.

In this paper, we propose a novel and effective data dimension reduction algorithm for solving the data overload problem in bioinformatics. Our work adopts the effective properties of heuristic feature selection to improve feature selection performance. This research introduces Artificial Bee Colony (ABC), which is a powerful swarm-based heuristic technique, as the feature selection method. ABC generally allows the solutions to converge more quickly than other heuristic methods such as GA and PSO. The selected features from ABC are evaluated for the fitness or quality of the feature subsets with the kNN method, which is the most simple and fundamental classifier. Our proposed method is applied to standard datasets for gene sequence analysis from [8–13]. We also explore the possibility of adopting our algorithm outside the scope of bioinformatics using a dataset in the context of autism behavior analysis.

This paper is organized as follows. A literature survey of related research is presented in Section 2. Section 3 provides a brief overview of methods used in this study including artificial bee colony and k-nearest neighbor algorithms. Section 4 describes our proposed method, ABC-kNN. Section 5 discusses the experimental results in gene sequence analysis. The use of our proposed method on real autism data is presented in Section 6. Finally, we discuss our conclusions in Section 7.

2. Related research in dimension reduction

In recent years, many research studies have adopted dimension reduction as a pre-analysis process in order to separate the irrelevant dimensions from the significant data. This process can be divided into feature selection and feature extraction. Feature selection is the method of selecting a feasible subset of features from the original set of candidate features. Feature extraction is a method that is used to extract meaningful features from the original data in order to reduce the multidimensional data to a lower dimensional space. This section summarizes the concepts of both dimension reduction techniques and presents previous work related to each method.

2.1. Feature extraction

A feature extraction method is a process that attempts to discover the dominant characteristics or dimensions in unstructured data. The method will use some mathematical function to transform the high-dimensional input data into lower dimensional data which can be represented as a set of features. In terms of mathematics, the algorithms in this technique can be divided into two types based on the characteristics of the mapping function: linear and non-linear.

In linear feature extraction, the methods will discriminate the less desired features from the original features by using a linear transformation. Principal component analysis (PCA) is the most popular linear feature extraction method. The method will construct a low dimensional dataset with minimal loss of information from the original data by computing a reduced dimensional dataset that has the largest variance. This variance depends on the covariance matrix of data which can be estimated by using the concept of eigenproblem (related to the eigenvector and eigenvalues of the data). PCA has been successfully applied for solving many problems. For example, Yang et al. [14] proposed PCA as a novel feature extraction method in an image recognition problem, using so called two-dimensional principal component analysis. In this work, the covariance matrix of PCA can be constructed directly by using the original image matrices; thus, the method requires less computation time than PCA but can provide higher recognition accuracy. Park et al. [15] combined the scheme of PCA feature extraction with the class label of each data in order to improve the effectiveness of image classification. Their results suggest that PCA can provide minimum error and a reduced number of features.

Similar to PCA, Etamad and Chellapa [16] proposed a linear feature extraction method called linear discriminant analysis (LDA). LDA attempts to identify a small set of relevant data features that can group patterns of the same class while separating them from the other classes. The method uses the concept of eigenproblem to produce an optimal linear transformation as does the PCA method. However, LDA is a supervised dimension reduction method so it requires the class information of data in order to optimize the line or plane separating the data in different classes. This work demonstrates the effective performance of LDA method as the results can be generated in real time when LDA is performed in applicable face-recognition tasks.

Another well-known linear feature extraction is presented in [17] which is called the independent component analysis (ICA). Unlike PCA and LDA, the discriminating process of ICA for reducing irrelevant features is done by statistical analysis. ICA will search for low dimensional projections of the data whose probability distributions are statistically independent and have non-Gaussian distributions.

Although PCA and other linear methods succeed as dimension reduction methods for linear data, they are not effective for data with non-linear distributions. For such problems, a non-linear feature extraction method must be applied. The concept of kernel-based approach and the manifold learning algorithms are often adapted in several non-linear feature extraction methods. The kernel-based approach uses an implicit function to transform the non-linear data into a feature space with high dimensionality.