Chapter 4

Nature Inspired Feature Selector for Effective Data Classification in Big Data Frameworks

Appavu Alias Balamurugan Subramanian
K.L.N. College of Information Technology, India

ABSTRACT

In high dimensional space finding clusters of data objects is challenging due to the curse of dimensionality. When the dimensionality increases, data in the irrelevant dimensions may produce much noise. And also, time complexity is the major issues in existing approach. In order to rectify these issues our proposed method made use of efficient feature subset selection in high dimensional data. We are considering the input dataset is the high dimensional micro array dataset. Initially, we have to select the optimal features so that our proposed technique employed Social Spider Optimization (SSO) algorithm. Here the traditional Social Spider Optimization is modified with the help of fruit fly optimization algorithm. Next the selected features are the input for the classifier. The classification is performed using optimized radial basis function based neural network (ORBFNN) technique to classify the micro array data as normal or abnormal data. The effectiveness of RBFNN is optimized by means of artificial bee colony algorithm (ABC).

1. INTRODUCTION

The profound increase in the technologies and their usage has led to the huge amount of data in terms of both number of attributes or features as well as records. Hypothetically, it appears to be coherent that more number of features means more precise information, nevertheless increase in number of features has created the curse of dimensionality problem. This implies that with the escalation in the number of dimensions, the performance of conventional algorithms begins to degrade. The nearness of pertinent and unessential features in the data can impede the working of classifier. They can make the classifier computationally more expensive and can also generate overfitted data model. In order to solve this problem, dimensionality reduction via feature selection is an appropriate strategy. Feature selection comprises of

DOI: 10.4018/978-1-5225-5852-1.ch004

choosing the pertinent features and disposing the insignificant ones to get a subset of features that best represent the data without losing any information. In addition, reduction in number of features helps in reducing the computational cost. Feature selection is an active area of research most applied in pattern recognition, information mining, genomics, bioinformatics, and computational science due to its added advantages.

These feature selection techniques are also known as feature subset selection strategies or feature ranking strategies. Feature Ranking (FR) strategies positions the features as indicated by their value in the classification problem. The Feature Subset Selection (FSS) strategies focus on providing the set of most important features for the classification model. Feature selection procedures are also classified as filters, wrappers, and embedded strategies. In wrapper techniques a subset of features is used to train the model iteratively. Subsequently, based on the performance of the model, features are added or removed features from the subsets. On the other hand, filter methods are commonly used as a preprocessing step. It is independent of the classification algorithm where features are chosen on the basis of their correlation with the outcome variable obtained by performing varied statistical tests. A hybrid feature selection technique known as embedded methods combines the qualities' of both the filter and wrapper techniques. These methods are implemented using the algorithms having their feature selection methods like LASSO and RIDGE regression which have inbuilt mechanism to control overfitting.

Researchers have characterized feature subset selection (FSS) as a search problem. In order to obtain the ideal subset of features, the most mainstream variable selection techniques for the most part incorporate forward, in reverse and floating successive strategies, which dependably utilize heuristic ways to deal with given an imperfect arrangement. The chosen ideal subsets of features produce higher classification precisions as compared to original set of features. After feature selection, two-stage classification procedure is applied to training and test data to compute the classification accuracy of the proposed approach. Rest of the chapter is organized as follows: Section 2 presents literature survey followed by proposed methodology in section 3. Experiments and results are discussed in Section 4 followed by conclusion.

2. LITERATURE SURVEY

Suhail Khokhar et al. (2016) proposed a novel method of programmed classification of single and cross breed PQDs. The proposed algorithm comprised of the Discrete Wavelet Transform (DWT) and Probabilistic Neural Network based Artificial Bee Colony (PNN-ABC) ideal feature selection of PQDs. DWT with Multi-Resolution Analysis (MRA) was utilized for the feature extraction of the noise. The PNN classifier was utilized as a successful classifier for the classification of the PQDs. It was observed that the new PNN-ABC based feature selection approach was more efficient for classification of PQDs.

Qi *et al.* (2016) emphasized that feature selection and classification have an important role in the field of Hyper Spectral Image (HSI) investigation They addressed the issue of HSI classification from the three perspectives. Primarily, they exhibited a novel basis by standard deviation, Kullback–Leibler separation and connection coefficient for feature selection. Subsequently they improved the SVM classifier by exploring the e most proper estimation of the parameters utilizing particle swarm optimization (PSO) with transformation instrument. At long last, they proposed a group learning system, which applied the boosting strategy to take in various kernel classifiers for classification issues. Trials were led on benchmark HSI classification information sets. The assessment comes about demonstrated that the proposed strategy could accomplish preferred precision and proficiency over best in class strategies.

Related Content

A Smart Helmet Framework Based on Visual-Inertial SLAM and Multi-Sensor Fusion to Improve Situational Awareness and Reduce Hazards in Mountaineering

Charles Shi Tan (2023). *International Journal of Software Science and Computational Intelligence (pp. 1-19).*

www.irma-international.org/article/a-smart-helmet-framework-based-on-visual-inertial-slam-and-multi-sensor-fusion-to-improve-situational-awareness-and-reduce-hazards-in-mountaineering/333628

Intra-Class Threshold Generation in Multimodal Biometric Systems by Set Estimation Technique

Dhiman Karmakar, Madhura Dattaand C.A. Murthy (2013). *International Journal of Software Science and Computational Intelligence (pp. 22-32).*

www.irma-international.org/article/intra-class-threshold-generation-in-multimodal-biometric-systems-by-set-estimation-technique/103352

Intelligent Fault Recognition and Diagnosis for Rotating Machines using Neural Networks

Cyprian F. Ngolah, Ed Mordenand Yingxu Wang (2011). *International Journal of Software Science and Computational Intelligence (pp. 67-83).*

www.irma-international.org/article/intelligent-fault-recognition-diagnosis-rotating/64180

Conservation of Information (COI): Geospatial and Operational Developments in E-Health and Telemedicine for Virtual and Rural Communities

Max E. Stachura, Elena V. Astapova, Hui-Lien Tung, Donald A. Sofge, James Grayson, Margo Bergmanand Joseph Wood (2012). *Machine Learning: Concepts, Methodologies, Tools and Applications (pp. 1146-1167).*

www.irma-international.org/chapter/conservation-information-coi/56191

Chaotic Tornadogenesis Optimization Algorithm for Data Clustering Problems

Ravi Kumar Saidalaand Nagaraju Devarakonda (2018). *International Journal of Software Science and Computational Intelligence (pp. 38-64).*

www.irma-international.org/article/chaotic-tornadogenesis-optimization-algorithm-for-data-clustering-problems/199016