Expert Systems With Applications 56 (2016) 28–47

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Velocity Bounded Boolean Particle Swarm Optimization for improved feature selection in liver and kidney disease diagnosis

S. Gunasundari[a,b,*], S. Janakiraman[a], S. Meenambal[c]

[a] Pondicherry University, India
[b] Velammal Engineering College, India
[c] Madurai Medical College, India

ARTICLE INFO

Article history:
Received 15 November 2015
Revised 22 February 2016
Accepted 23 February 2016
Available online 2 March 2016

Keywords:
Boolean Particle Swarm Optimization
BPSO
Feature selection
Liver cancer
Kidney cancer

ABSTRACT

Cancer is one of the foremost causes of death and can be reduced by early diagnosis. Computer Aided Diagnostic system plays an important role in the detection of cancer. Feature selection is an important preprocessing step in the classification phase of the diagnostic system. The feature selection is a NP – hard challenging problem that have many applications in the area relevant to expert and intelligent system. In this study, two new modified Boolean Particle Swarm Optimization algorithms are proposed namely Velocity Bounded BoPSO (VbBoPSO) and Improved Velocity Bounded BoPSO (IVbBoPSO) to solve feature selection problem. Compared to the basic Boolean PSO, these improved algorithms introduce V_{min} parameter that makes it more effective in solving feature selection problem. The performance of VbBoPSO and IVbBoPSO are tested over 28 benchmark functions provided by CEC 2013 session. A comparative study of proposed algorithms with the recent modification of Binary Particle Swarm Optimization and Boolean PSO (BoPSO) is provided. The results prove that the proposed algorithms improve the performance of BoPSO significantly. In addition, the proposed algorithms are tested in the feature selection phase of intelligent disease diagnostic system. Experiments are carried out to select elite features from the liver and kidney cancer data. Empirical results illustrate that the proposed system is superior in selecting elite features to achieve highest classification accuracy.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Primary liver cancer is the sixth most frequent cancer globally. Liver cancer is the second leading cause of cancer death (Stewart & Wild, 2014, chap. 1.1). Radiographic imaging modalities like Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) help in liver cancer diagnosis. Often, CT is used for identifying different diseases including cancer, hemangioma, hepatocellular adenoma etc., since the image permits a physician to verify the presence of a lesion and to measure its size. A benign tumor such as hemangioma, hepatocellular adenoma, and Focal Nodular Hyperplasia (FNH) are not cancer cells but malignant tumors such as Hepatic Cellular Carcinoma (HCC) and cholangiocarcinoma are cancer cells. Malignant tumor (cancer) has a similar visual feature with the benign tumor. There is a need to differentiate benign and malignant tumors, as treatments are different for both. The human observations are not 100% reliable in diagnosis and the biopsy is an invasive method (Kumar, Moni, & Rajeesh, 2013). As the effort of the physician is more challenging, expert system has become essential to classify tumor. With the advent of artificial intelligence, above-mentioned problem can be resolved. Computer Aided Diagnostic system (CAD) is an expert system, which is used to support doctors with diagnostic advice to categorize diseases accurately. The CAD systems follow a structure and it is a vital tool for health professionals.

From the literature, it is seen that researchers have worked for the development of CAD system. CAD system has been developed recently to assist doctors to identify liver diseases accurately and it is surveyed (Gunasundari & Janakiraman, 2013a). Many authors have focused on diagnosis of liver diseases like HCC, hemangioma, and cirrhosis (Mala & Sadasivam, 2010; Gunasundari & SuganyaAnanthi, 2012; Gunasundari & Janakiraman, 2013b; Mougiakakou, Valavanis, Nikita, & Nikita, 2007; Kumar, Moni, & Rajeesh, 2012; Bharathi & Ganesan, 2008). However, only a few studies have been carried out to diagnose hepatic adenoma, cholangiocarcinoma, fatty liver, and FNH (Wu, Lee, & Chen, 2013; Mitrea, Nedevschi, Socaciu, & Badea, 2012). Hence, there is a necessity for a liver CAD system to distinguish the

* Corresponding author at: Pondicherry University, India. Tel.: +91 8754509510 (mobile).
E-mail addresses: gunapondyuniv@gmail.com, gunaselvaraj@yahoo.com (S. Gunasundari), jana3376@yahoo.co.in (S. Janakiraman), kuttimbbs.s@gmail.com (S. Meenambal).

http://dx.doi.org/10.1016/j.eswa.2016.02.042
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

benign tumor and malignant tumor considering the aforementioned diseases.

Renal Cell Carcinoma (RCC) is the seventh most common cancer in males and ninth most common in females in UK (Haynes, 2013). There is an assessment that quarter of a million people in USA are living with kidney cancer. The number of kidney cancer patients has been increasing to 51,000 every year (Linguraru et al., 2009). The kidney diseases like angioyolipama and oncocytoma are benign tumor and, RCC and Transitional Cell Carcinoma (TCC) are malignant tumors. Kidney CAD systems are developed recently to assist doctors to identify kidney diseases accurately (Linguraru et al., 2009; Juan et al., 2010; Gowrishankar et al., 2015). Only limited authors have been working on kidney diseases such as TCC and angioyolipama. Therefore, there is a requirement for a kidney CAD system to characterize the benign tumor from the malignant tumor, which considers the rare diseases.

Classification of benign and malignant tumor based on shape or gray level information are not feasible. However, it is feasible to use texture, since homogeneous texture is maintained along with series of slices (Kumar et al., 2013). Therefore, in CAD system, for identifying liver and kidney diseases, suspicious region from an abdominal Computed Tomography necessitates texture differentiation. Among the general texture descriptors, Spatial Gray-Level Dependence Matrix (SGLDM) is superior for extracting the textural properties of blocks defined in the spatial domain of image data (Mala & Sadasivam, 2010; Wu et al., 2013). CAD system use learned models to differentiate tumors. To construct learned models textural features are required. Not all the features are important for the classification. A number of irrelevant features degrade the generalized performance of the classifier. Classifier needs feature selection (FS) procedure to improve the performance of the classifier thereby increasing the quality of the dataset by removing spurious, irrelevant, and redundant features. FS spreads all the way through many fields like data mining and pattern recognition. It is not well known in advance, which feature subset is best for classification. An exhaustive search considering all possible subsets is practically hard in most circumstances. Manual selection of dominant features is not consistent and it is not possible. Hence, it mandates a computational method for FS. Searching for the optimal subset, which categorizes the disease accurately, is a quite challenging task.

Literature reveals that, the methods for FS are categorized into filter, wrapper, and hybrid. Filter methods require statistical analysis of features to select the features whereas wrapper methods need learning model to select the features. Benefits of filter methods are less computationally expensive, scalability and independence from the classifier (Saeys, Inza, & Larranaga, 2007). Compared to filter method, wrapper methods do well because the selected subset is optimized for the classification method. Wrapper methods outperform filter methods (Wu et al., 2013; Chuang, Tsai, & Yang, 2011a; Al-Ani, Alsukker, & Khushaba, 2013). Wrappers usually achieve better classification rates than filters since they adjust to the precise interactions between the classifier and the data set. It has the ability to generalize, because they typically use k-fold cross-validation procedures of predictive accuracy. Hybrid methods utilize both approaches to select features. Best feature subsets are identified using different ways like sequential forward search (Guan, Liu, & Qi, 2004), sequential backward search (Gasca, Sanchez, & Alonso, 2006), bidirectional selection (Caruana & Freitag, 1994), and complete search (Liu & Tu, 2004). Sequential methods are easy to implement but it is affected by nesting effect. This problem is solved using "plus-*l*-take away-*r*" procedure that does one time's forward selection trailed by *r* time's backward removal (Stearns, 2003). However, it is tough to fix the optimal values of (*l, r*). Afore-mentioned methods use local search instead of global search. Consequently, these algorithms are complex to find optimal solutions. Thus, researchers use Meta heuristic algorithms, which perform global search to find the high quality solutions.

Many meta-heuristic approaches such as Genetic Algorithm (GA) (Oh, Lee, & Moon, 2004), Simulated Annealing (Lin, Tseng, Chou, & Chen, 2008), Tabu Search (Zhang & Sun, 2002), Ant Colony Optimization (Kanan & Faez, 2008), and Particle Swarm Optimization (PSO) (Chuang et al., 2011a) have also been recently applied for FS. Among the wrapper meta-heuristic methods, GA and PSO are two popular evolutionary algorithms (Wu et al., 2013). Despite GAs have the ability to reach near-optimal solutions for large problems, it requires a long processing time to reach the near-optimal solution (Chuang et al., 2011a). PSO has been applied to many domains due to its simplicity, fewer adjustable parameters, and cost effectiveness. In this study, PSO, which is very proficient to look for large solution spaces, has been investigated. Recently, variants of BPSO have become popular and it is applied to various optimization problems successfully (Shen, Jiang, Jiao, Shen, & Yu, 2004; Wang, Wang, Zhen, & Zhen, 2008). However, BPSO is not widely researched like continuous PSO.

Binary Particle Swarm Optimization (BPSO) is applied to many UCI data sets for solving FS problem. It performs well on certain data sets whereas, for some data sets such as Flag, Connect, Vehicle, and German, classification accuracy is lower than 80% (Chuang et al., 2011a; Chuang, Yang, & Li, 2011b; Yuanning et al., 2011). The main limitation is the lack of exploration ability. Variants of BPSO have been used for feature selection in many disease diagnostic systems (Sahu & Mishra, 2012; Babaoglu, Findik, & Ulker, 2010; Chang, Lai, Lai, & Chen, 2013). But, it could not identify the global optima for many standard UCI medical data sets such as Pima Indian Diabetics (accuracy – 69%), Arrhythmia (accuracy – 75.4%), 9 Tumor (accuracy – 85%) and 14 Tumor (accuracy – 70%). The main defect in the Binary PSO is that it gets entrapped in the local optima always during the later stages of iterations (Luh & Lin, 2011) and it has poor scaling behavior (Gorse, 2013).

To overcome the shortcoming of BPSO, a novel BPSO called BoPSO was introduced in the year 2005. BoPSO works with binary variables and operators. BoPSO is more suitable for discrete problems. Nevertheless it is not always effective in higher dimensional spaces (Gorse, 2013).To improve the performance of BoPSO in high dimensional search space, this study proposes Velocity Bounded Boolean Particle Swarm Optimization (VbBoPSO). To avoid stagnation in the subsequent iteration, a novel variant of VbBoPSO (IVbBoPSO) is introduced. Both VbBoPSO and IVbBoPSO are tested with CEC 2013 benchmark functions and in FS process of liver and kidney CAD system. The following sections provide a more detailed description of the proposed algorithm. The rest of this paper is organized as follows. Section 2 discusses the related work briefly. Section 3 reviews the BoPSO algorithm. Section 4 describes in detail about the proposed algorithm and structure of CAD system. Section 5 reports the experimental design and results of VbBoPSO. Finally, Section 6 concludes the work by giving summarization and reports with future improvements.

2. Related work

Classification is one of the most considered problems in machine learning and data mining. Due to plenty of noisy, irrelevant, or misleading features, in real world problems, FS has become one of the most significant preprocessing tools for classification. Evolutionary computing can be applied to FS problems where traditional methods are hard to apply. There exists BPSO, which is a binary version of PSO that is widely used for selecting the best features.

A FS algorithm that integrates the particle swarm optimization and neural network with the Boltzmann function is implemented to select significant features for the classification of lymph nodes

Download English Version:

https://daneshyari.com/en/article/382318

Download Persian Version:

https://daneshyari.com/article/382318

Daneshyari.com