Expert Systems with Applications 37 (2010) 7648–7655

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Chest diseases diagnosis using artificial neural networks

Orhan Er [a,b], Nejat Yumusak [c], Feyzullah Temurtas [a,*]

[a] Bozok University, Department of Electrical and Electronics Engineering, 66200 Yozgat, Turkey
[b] Sakarya University, Department of Electrical and Electronics Engineering, 54187 Adapazari, Turkey
[c] Sakarya University, Department of Computer Engineering, 54187 Adapazari, Turkey

ARTICLE INFO

Keywords:
Chest disease diagnosis
Multilayer neural network
Probabilistic neural network
Learning vector quantization neural network
Generalized regression neural network
Radial basis function neural network

ABSTRACT

Chronic obstructive pulmonary, pneumonia, asthma, tuberculosis, lung cancer diseases are the most important chest diseases. These chest diseases are important health problems in the world. In this study, a comparative chest diseases diagnosis was realized by using multilayer, probabilistic, learning vector quantization, and generalized regression neural networks. The chest diseases dataset were prepared by using patient's epicrisis reports from a chest diseases hospital's database.

1. Introduction

The chest contains the main respiration and circulation organs which sustain some of the most critical life functions of the body. Millions of people are diagnosed every year with a chest disease in the world. Tuberculosis (TB), chronic obstructive pulmonary disease (COPD), pneumonia, asthma, lung cancer diseases are the most important chest diseases which are very common illnesses in the world (MedHelp: http://www.medhelp.org/Medical-Dictionary/Terms/2/8964.htm (accessed 18.03.09)).

TB is an infectious disease, caused in most cases by microorganisms called *Mycobacterium tuberculosis*. The microorganisms usually enter the body by inhalation through the lungs. They spread from the initial location in the lungs to other parts of the body via the blood stream, the lymphatic system, via the airways or by direct extension to other organs. TB is a major cause of illness and death worldwide and globally, 9.2 million new cases and 1.7 million deaths from tuberculosis occurred in 2006 (Enarson, Rieder, Arnadottir, & Trébucq, 2000; Royal College of Physicians of London, 2006; World Health Organization, 2008).

COPD is a disease state characterized by airflow limitation that is not fully reversible. The airflow limitation is usually both progressive and associated with an abnormal inflammatory response of the lungs to noxious particles or gases (Celli & MacNee, 2004). Clinically, patients with COPD experience shortness of breath (dyspnea) and cough, productive of an excess of mucus. There may also be wheeze (Jeffery, 1998). According to the World Health Organization (WHO) data is found 600 million patients who have COPD and every year 2.3 million persons die because of COPD in the world (Sönmez & Uzaslan, 2006).

Pneumonia is an inflammation or infection of the lungs most commonly caused by a bacteria or virus. Pneumonia can also be caused by inhaling vomit or other foreign substances. In all cases, the lungs' air sacs fill with pus, mucous, and other liquids and cannot function properly. This means oxygen cannot reach the blood and the cells of the body effectively. According to the World Health Organization (WHO) data, every year approximate 2.4 million persons die because of pneumonia (Global Action Plan for the Prevention, 2007).

Asthma is a chronic disease characterized by recurrent attacks of breathlessness and wheezing. During an asthma attack, the lining of the bronchial tubes swell, causing the airways to narrow and reducing the flow of air into and out of the lungs. Recurrent asthma symptoms frequently cause sleeplessness, daytime fatigue, reduced activity levels and school and work absenteeism. Asthma has a relatively low fatality rate compared to other chronic diseases. WHO estimates that 300 million people currently suffer from asthma. Asthma is the most common chronic disease among children (http://www.who.int/en/ (accessed 18.03.09)).

Lung cancer is a disease of uncontrolled cell growth in tissues of the lung. This growth may lead to metastasis, which is the invasion of adjacent tissue and infiltration beyond the lungs. The vast majority of primary lung cancers are carcinomas of the lung, derived from epithelial cells. Lung cancer, the most common cause of cancer-related death in men and the second most common in women, is responsible for 1.3 million deaths worldwide annually (http://www.who.int/en/ (accessed 18.03.09)).

Artificial neural network (ANN) structures for classification systems in medical diagnosis are increasing gradually. The multilayer neural network (MLNN), probabilistic neural network

* Corresponding author. Tel.: +90 532 207 56 33.
E-mail address: temurtas@gmail.com (F. Temurtas).

0957-4174/$ - see front matter
doi:10.1016/j.eswa.2010.04.078

(PNN), learning vector quantization (LVQ) neural network, generalized regression neural network (GRNN), and radial basis function (RBF) neural network structures have been successfully used in replacing conventional pattern recognition methods for the disease diagnosis systems. As for other clinical diagnosis problems, the neural network classification systems have been used for chest diseases diagnosis problem also. There have been several studies reported focusing on chest diseases diagnosis using artificial neural network structures (Aliferis, Hardin, & Massion, 2002; Ashizawa et al., 2005; Coppini, Miniati, Paterni, Monti, & Ferdeghini, 2007; El-Solh, Hsiao, Goodnough, Serghani, & Grant, 1999; Er, Sertkaya, Temurtas, & Tanrikulu, 2009; Er & Temurtas, 2008; Er, Temurtas, & Tanrikulu, 2010; Hanif, Lan, Daud, & Ahmad, 2009; Paul, Ben, Thomas, & Robert, 2004; dos Santos, Pereira, & de Seixas, 2004; Temurtas, 2009). These studies have applied different neural networks structures to the various chest diseases diagnosis problem and achieved high classification accuracies using their various dataset.

The multilayer neural network structure are the most common neural network structure which have been successfully used for the disease diagnosis systems (Delen et al.,2005; Er & Temurtas, 2008; Kayaer & Yıldırım, 2003; Temurtas, 2009). The back-propagation (BP) algorithm (Rumelhart, Hinton, & Williams, 1986) is widely recognized as a powerful tool for training of the MLNN structures. However, BP algorithm suffers from a slow convergence rate and often yields suboptimal solutions (Brent, 1991; Gori & Tesi, 1992). A variety of related algorithms have been introduced to address that problem and a number of researchers have carried out comparative studies of MLNN training algorithms (Gulbag & Temurtas, 2006; Hagan, Demuth, & Beale, 1996; Hagan & Menhaj, 1994). Levenberg–Marquardt (LM) algorithm (Hagan & Menhaj, 1994) used in this study provides generally faster convergence and better estimation results than other training algorithms (Er & Temurtas, 2008; Gulbag & Temurtas, 2006).

The probabilistic neural network structures provide a general solution to pattern classification problems by following an approach developed in statistics, called Bayesian classifiers. The PNN uses a supervised training set to develop distribution functions within a pattern layer. Training of the PNN is much simpler than that of the MLNN. However, the pattern layer can be quite huge if the distinction between categories is varied and at the same time quite similar in special areas (Speckt, 1990). Because the PNN provides a general solution to pattern classification problems, it is suitable for the disease diagnosis systems (Er, Sertkaya, et al., 2009; Temurtas, 2009; Temurtas, Yumusak, & Temurtas, 2009).

The classification of the learning vector quantization neural network structure is based on the similarity of the unknown data and these prototypes. An LVQ neural network has a competitive layer and linear output layer. The competitive layer learns to classify input vectors. The linear output layer transforms the competitive layer's classes into target classifications defined by the user. The classes learned by the competitive layer can be referred as subclasses and the classes of the linear output layer can be referred as target classes (Kohonen, 1990; Kohonen, 1997; Matlab Documentation, 2004). The LVQ network structures have been successfully used for the disease diagnosis systems (Er, Sertkaya, et al., 2009; Er, Temurtas, et al., 2010; Temurtas, 2009).

Like probabilistic neural networks, generalized regression neural networks are known for their ability to train in only one pass of the training set using sparse data sets. Rather than categorizing data like PNN, however, GRNN applications are able to produce continuous valued outputs. GRNN is especially useful for continuous function approximation, and can fit multidimensional surfaces through data (Speckt, 1991). There has been some studies reported focusing on the chest disease diagnosis using GRNN structure (El-Solh et al., 1999).

The radial basis functions greatly reduce the training time and make related analyses much easier. The RBF neural networks are quite suitable for implementing multi-class and high-dimensional classification problems (Chen & Lin, 1993; Daqi, Shuyan, & Yan, 2004). The RBF network structures have been successfully used for the disease diagnosis problems also (Hanif et al., 2009).

In this study, a comparative chest diseases diagnosis was realized by using multilayer, probabilistic, learning vector quantization, generalized regression, and radial basis function neural networks. The chest diseases dataset were prepared by using patient's epicrisis reports from a chest diseases hospital's database. The study aims also to provide machine learning based decision support system for contributing to the doctors in their diagnosis decisions.

2. Method

2.1. Data source

In order to perform the research reported in this article, the patient's epicrisis taken from Diyarbakir Chest Diseases Hospital from southeast of Turkey was used. The dataset were prepared using these epicrisis reports. The dataset which consists of the chest disease measurements contains six classes and 357 samples. The class distribution is

- Class 1: Tuberculosis (50).
- Class 2: COPD (71).
- Class 3: Pneumonia (60).
- Class 4: Asthma (44).
- Class 5: Lung Cancer (32).
- Class 6: Normal (100).

All samples have thirty eight features. These features are (Laboratory examination): complaint of cough, body temperature, ache on chest, weakness, dyspnoea on exertion, rattle in chest, pressure on chest, sputum, sound on respiratory tract, habit of cigarette, leucocyte (WBC), erythrocyte (RBC), trombosit (PLT), hematocrit (HCT), hemoglobin (HGB), albumin2, alkalen phosphatase 2 L, alanin aminotransferase (ALT), amylase, aspartat aminotransferase (AST), bilirubin (total + direct), CK/creatine kinase total, CK–MB, iron (SERUM), gamma–glutamil transferase (GGT), glukoz, HDL cholesterol, calcium (CA), blood urea nitrogen (BUN), chlorine (CL), cholesterol, creatinin, lactic dehydrogenase (LDH), potassium (K), sodium (NA), total protein, triglesid, uric acid.

2.2. Previous studies

There have been several studies reported focusing on chest disease diagnosis problem using artificial neural network structures as for other clinical diagnosis problems. These studies have applied different neural networks structures to the various chest diseases diagnosis problem using their various dataset (Aliferis et al., 2002; Ashizawa et al., 2005; Coppini et al., 2007; El-Solh et al., 1999; Er & Temurtas, 2008; Er, Temurtas, et al., 2010; Hanif et al., 2009; Paul et al., 2004; dos Santos et al., 2004).

El-Solh et al. used a generalized regression neural network (GRNN) using clinical and radiographic information to predict active pulmonary tuberculosis at the time of presentation at a health-care facility that is superior to physicians' opinion (El-Solh et al., 1999). The input patterns were formed by 21 distinct parameters which were divided into three groups: demographic variables, constitutional symptoms, and radiographic findings. The output of the GRNN provided an estimate of the likelihood of active pulmonary tuberculosis. The authors utilized