Novel Computer-Aided Digital Data Mining Tools For Early Breast Cancer Detection and Diagnosis For Lesion Classification

Dr. Hamparsum Bozdoğan [a]

[a] *Toby and Brenda McKenzie Chair Professor, Department of Statistics, Operations and Management Science and Center for Intelligent Systems and Machine Learning (CISML) The University of Tennessee Knoxville, TN, U.S.A.*

Abstract: *Breast cancer is the second-leading cause of death among women worldwide, causing the death of nearly half a million women every year. Radiologists still miss up to 30 percent of breast lesions in mammograms. What can data mining do?*

In this talk, we present novel data mining techniques for computer-aided detection (CAD) of breast cancer by introducing and developing two flexible supervised and unsupervised classification methods. For the supervised classification method, we develop what is called Probabilistic Kernel Quadratic Discriminant Analysis (PKQDA), and for the unsupervised classification case we develop mixture-model cluster analysis (MMC) under different covariance structures. In using both methods we strive to classify the signs of disease tissues on the resulting digital radiographic images (i.e., mammograms) in order to help radiologists to reach diagnostic decisions as a second eye. Mammography screening programs have been adopted worldwide to look for possible signs of breast cancer on asymptomatic patients at an early stage, especially when the chance of survival is highest.

Mixture-model cluster analysis is an unsupervised classification model that learns the actual number of clusters without knowing a priori the classification of cancerous lesions or labels. Both approaches use a model selection criterion based on the information-theoretic measure of complexity (ICOMP) index introduced by this author, which allows robust statistical inference to detect cancerous lesion classification and diagnosis. An experimental case study demonstration of both methods is presented by conducting a detailed analysis of a real data set on two breast cancer groups ("Benign"/"Malignant") composed of n = 1269 Italian patients with p = 132 continuous features. The efficiency and robustness of our two approaches are presented and compared with results obtained by using the Support Vector Machines (SVMs) method approach commonly employed in Computer System Detection (CAD) of breast tumors. It is shown that our two methods constitute a new and novel approach.

Our results elucidate the current inferential problems often found in classical statistical data mining as a first step toward the specification of a robust classification model for breast cancer detection through image modeling.

The two proposed approaches have many other applications, not only in biomedical and health informatics, but also in a variety of business applications (e.g., detecting potential fraud and bankruptcy, performing customer profiling and market segmentation, auditing of accounting practices, and detecting potential threats).

Key Words: Breast cancer detection; Probabilistic Kernel Quadratic Discriminant Analysis; Mixture-model cluster analysis; Covariance Structures; Bayesian classification; Information Complexity

Sorumlu Yazarın Adresi

Dr. Hamparsum Bozdoğan, Email: bozdogan@utk.edu.