CHAPTER II

LITERATURE REVIEW

2.1 Review of Literatures Related to Statistical Inferences of Missing Data

T. Schneider [1] applied the expectation maximization (EM) algorithm, an iterative method, for the estimation of mean values and covariance matrices from incomplete datasets and for the imputation of missing values. Estimating the mean and the covariance matrix of an incomplete dataset and filling in missing values with imputed values was a nonlinear problem, which had to be solved iteratively. He proposed the regularized EM algorithm that was applicable to set of climate data, in which the number of variables typically exceeds the sample size. Because of the availability of climatic measurements was various space and time that set of climate data were usually incomplete. A test of the regularized EM algorithm with simulated surface temperature data demonstrates that the algorithm was applicable to typical set of climate data and lead to more accurate estimates of missing values than a conventional noniterative imputation technique.

G. McLachlan and T. Krishnan [2] suggested the situations where the EM can be applied include not only incomplete-data situations, where there are missing data, truncated distributions, grouped observation, but also a whole variety of situation where the completeness of the data is not all that natural or evident, and statistical model. The EM algorithm has found applications in almost all statistical contexts and in almost all fields where statistical techniques have been applied.

A. P. Dempster, N. M. Laird, and D. B. Rubin [3] proposed the Expectation-Maximization (EM) algorithm which is useful in a variety of incomplete-data problem. The EM algorithm is a largely applicable approach to the iterative computation of maximum likelihood (ML) estimates. There are two steps called the expectation step or the E-Step and the maximization step or the M-step. The situations where the EM algorithm is usefully applied can be described as incomplete-data problems, where ML estimation is made difficult by the lack of some part of data in a more familiar and simpler data structure. The EM algorithm is closely related to the ad hoc approach to estimating missing data, where the parameters are estimated after filling in initial values for the missing data. The latter are then updated by their predicted values using these initial parameter estimates. The parameters are then re-estimated, and so on, proceeded iteratively until convergence.

R. J. Hathway and J. C. Bezdek [4] proposed a new algorithm for doing fuzzy c-means (FCM) clustering of incomplete data sets. Numerical convergence properties of the new algorithm were discussed and all approaches were tested using real and artificially generated incomplete data sets. It was important for general statistical method to handle incomplete data based on the expectation maximization (EM) algorithm [3]. In brief, this approach iteratively produced maximum likelihood estimates using two-part iteration. First, (E-step) used current distributional parameter estimates to calculate expected values for the missing features. Second, (M-step) used the current expected values to complete the value data and then calculated improved (complete data) maximum-likelihood parameter estimates. These two parts, EM iteration continued using the improve parameter estimates until convergence was achieved.

2.2 Review of Literatures Related to Neural Inferences of Missing Data

C. E. Pedreira and E. Parente [5] presented a framework based on maximum likelihood density estimation for learning from high-dimensional data sets with arbitrary patterns of missing data. Learning in this framework was a classical estimation problem requiring an explicit probabilistic model and an algorithm for estimating the parameters of the model. Mixture models combined much of the flexibility of nonparametric methods with certain of the analytic advantages of parametric methods. This density-based approach was applicable to both supervised and unsupervised learning. This happened because the problem of estimating mixture densities could be viewed as a missing data problem (the label for the component densities were missing) and the Expectation-Maximization (EM) algorithm [3] could be developed to handle both kinds of missing data. Results from a classification benchmark which was presented.

C. E. Pedreira and E. Parente [6] introduced a new theoretical approach that enabled to deal with missing values attributes as inputs. The neural network inputs were now treated as random variables, and possibly with large variances in the case when the input variable information was not complete. Feedforward Neural Networks trained with backpropagation algorithm that had been successfully using in a variety of applications. This paper proposed a stochastic framework to backpropagation. In many practical situations, the input patterns had intrinsic doses of uncertainty. An automatic diagnosis system was used to illustrate this new technique potential.

A. Verikas, A. Gelzinis, K. Malmqvist, and M. Bacauskiene [7] proposed an approach using both labeled and unlabelled data to train a multilayer perceptron. The unlabelled data were iteratively pre-processed by a perceptron being trained to obtain the soft class label estimates. Because of one reason, conventional supervised learning approaches, the error back propagation had no direct way to incorporate unlabelled data, and discarded them.

M. P. Perrone and L. N. Cooper [8] presented a general theoretical framework for ensemble methods of constructing significantly improved regression estimates. They constructed a hybrid estimator which was better in the MSE sense than any estimator. The ensemble method had several properties: 1) it efficiently used all the networks. 2) It efficiently used all data for training without over-fitting. 3) It performed regularization by smoothing in functional space. 4) It utilized local minima to construct improved estimates. 5) It was suite for parallel computation. 6) It led to a very useful and natural measure of the number of distinct estimators. Hybrid neural network systems had been employed to improve results in classification and regression problems. The authors addressed the issues of optimal combination and efficient data usage in the framework of ensemble averaging.

M. C. Mozer [9] presented a general taxonomy of neural net architectures for processing time-varying patterns. The architecture based on a characterization of short term memory models along the dimensions of form, content and adaptability. Experiment on predicting future values of a financial time series were presented using several alternative memory models. The result served as a based line against which more sophisticated architecture could be compared.

2.3 Review of Literatures Related to Radar Rainfall Estimates

D. Rosenfeld, Wolff. D. B. and E. Amitai [10] introduced a simplified probability matching method that relies on matching the unconditional probabilities of gauge rain intensity (R) and effective radar intensity (Z_e). This was achieved by matching rain gauge intensities to radar reflectivity taken only from small "windows" centered about the gauges in time and space. The windows had to be small enough for the gauge to represent the rainfall depth within the radar window, and it had to be also large enough to encompass the timing and geometrical errors inherent to the observations. A relatively small sample size was required to achieve a stable $Z_e - R$ relation with the standard deviation of 15 % of R for a given Z_e. The Window Probability Matching Method (WPMM) significantly performed better rainfall integrations than the power law. The standard deviation of the WPMM rainfall integration, after correction for systematic bias errors, was only two-thirds of the standard deviation obtained when using power law based on disdrometer which measured drop size distribution. Author mentioned that the accuracy of the WPMM was provided upon its application to the data that had been objectively classified into different rain regimes.

D. Rosenfeld and E. Amitai [11] evaluated the accuracy of the estimation of $Z_e - R$ relationships for the Window Probability Matching Method (WPMM) and regression method. The evaluation was based on experiments of random sub-sampling of disdrometer obtained 1-min reflectivity Z_e and rain-rate R pairs. Geometrical mismatch and synchronization inaccuracies between the radar and rain gauges were simulated by desynchronization of dt minutes. The WPMM had significant advantage in estimating the rain intensities when geometrical and synchronization errors were introduced to the radar and rain gauge measured $Z_e - R$ pairs for simulating real-world radar and rain gauge comparisons. Regression-based relationships tended to overestimate the low rain intensities and underestimate the high rain intensities with the crossover at the estimated median rain volume intensity. Although, rain gauge bias correction might make the overall rain accumulation insensitive to the power of the $Z_e - R$ law, its appropriate selection had a major effect on the partition of rainfall amounts between weak and strong intensities or the partition between convective and stratiform rainfall.

Rongrui Xiao and V. Chandrasekar [12] introduced a neural network based approach to address radar rainfall estimation by taking into account the three-dimensional (3-D) structure of precipitation. A three-layer perceptron neural network was developed for rainfall estimation from radar measurements. The neural network was trained using the radar measurements as the input and the ground rain gage measurements as the target (output). The neural network based estimates were evaluated using data collected during the Convection and Precipitation Electrification (CaPE) experiment conducted over central Florida in 1991 [13]. The results of the evaluation showed that the neural network could be successfully applied to obtain rainfall estimates on the ground, based on radar observations. The rainfall estimates obtained from the neural network were shown to be better than those obtained from several existing techniques. The neural network based rainfall estimate offered an alternate approach to the rainfall estimation, and it could be implemented easily in operational weather radar systems.