XỬ LÝ DỮ LIỆU THIẾU TRONG NGHIÊN CỨU PHỤ TẢI BẰNG SUPPORT VECTOR REGRESSION (SVR)

DEALING WITH MISSING DATA FOR THE POWER LOAD STUDIES USING SUPPORT VECTOR REGRESSION (SVR)

Tác giả: *Nguyễn Tuấn Dũng, Nguyễn Thanh Phương*

Tổng Công ty Điện lực TP. Hồ Chí Minh; dungnt@hcmpc.com.vn
Trường Đại học Công nghệ TP. Hồ Chí Minh; nt.phuong@hutech.edu.vn

Tóm tắt:

Trong những năm gần đây, việc nghiên cứu và ứng dụng các kỹ thuật khai thác dữ liệu gặp phải nhiều khó khăn, thách thức lớn, trong đó có vấn đề giá trị thiếu, tức là có những giá trị thuộc tính của dữ liệu bị thiếu. Có nhiều nguyên nhân khác nhau dẫn tới hiện tượng này: thiết bị thu thập dữ liệu bị hỏng, sự thay đổi thiết kế thí nghiệm, sự từ chối cung cấp dữ liệu nhằm bảo vệ tính riêng tư, sự sơ suất khi nhập dữ liệu, các sự cố xảy ra trong quá trình truyền dữ liệu,…[1]. Trong đó, việc thiếu dữ liệu phục vụ công tác nghiên cứu, dự báo phụ tải điện là một trong những vấn đề nan giải đối với ngành điện. Hiện các công ty điện lực đang thực hiện việc này bằng cách nội suy từ các giá trị đo đếm của các ngày trước, giờ trước một cách thủ công, không chuẩn xác làm ảnh hưởng không nhỏ đến kết quả phân tích, xử lý dữ liệu trong quá trình nghiên cứu phụ tải. Bài báo đề xuất một phương pháp xử lý dữ liệu thiếu bằng cách xây dựng các mô hình hồi quy tối ưu hóa các thông số tự động thông qua quá trình huấn luyện học máy Support Vector Regression (SVR), từ đó ước lượng lại các dữ liệu đã mất hoặc không ghi nhận được trong quá trình đo đếm.

Từ khóa: Thiếu dữ liệu; Ước lượng; Số liệu đo đếm; Phụ tải điện; SVM; SVR.

Abstract:

In recent years, the research and the application of data mining techniques have encountered many difficulties and challenges, including the missing value problem i.e. the attribute values of the data are missing . There are many different causes of this phenomenon: damaged data collection equipment, the change of design of experiments, the refusal to provide the data in order to protect privacy, the mistake when importing data, the incident occurrence during the data transmission... [1]. In particular, the lack of data is one of the problems for the power sector. The power companies are doing this manually, causing influence on results analysis. This paper proposes a method of handling missing data by building the regression model to optimize parameters automatically through Support Vector Regression (SVR), machine learning training which estimates the lost data or unrecorded data during the measurement.

Key words: Missing data; Estimation; Measurement data; Power load; SVM; SVR.