

Bài 1 :

PHÂN TÍCH THÀNH PHẦN CHÍNH - PHƯƠNG PHÁP TRÍCH RÚT ĐẶC TRƯNG HỮU HIỆU TRONG KHÔNG GIAN CÓ SỐ CHIỀU LỚN

**PRINCIPAL COMPONENT ANALYSIS - AN EFFECTIVE FEATURE EXTRACTION
METHOD IN HIGH DIMENSION SPACE**

Nguyễn Thị Phương Nga

Đại học Công Nghiệp Hà Nội

Email: nga582002@yahoo.com

TÓM TẮT

Khi xử lý các bài toán trong thực tế ta thường gặp tập mẫu có số chiều mô tả lớn. Hơn nữa, trong tập mẫu có thể chứa các thuộc tính chứa ít thông tin hữu dụng, các thuộc tính tương tác lẫn nhau. Điều đó làm cho việc xử lý gặp rất nhiều khó khăn. Vấn đề đặt ra là phải trích rút được các đặc trưng từ tập thuộc tính mô tả tập dữ liệu mẫu đó, phục vụ cho việc phát hiện tri thức. Trong bài báo này chúng tôi chỉ ra rằng phân tích thành phần chính (PCA-Principal Component Analysis) là một phương pháp hiệu quả, thực hiện công việc này trong trường hợp không gian bài toán lớn. Kết quả thực nghiệm với bài toán dự báo hoạt tính sinh học dựa trên tập mẫu PDGFR (platelet derived growth factor receptor) Inhibitors đã chứng tỏ hiệu quả của phương pháp.

ABSTRACT

When solving maths problems, we often meet sample sets with very high dimension space. Moreover, in those sample sets there may be poor description parameters, mutually interactive parameters that makes problem resolution difficult. The matter is to extract features from description attributes in sample sets for the data mining process. This paper, show that Principal Component Analysis (PCA) is an effective method for feature extraction in high dimension space. Experiments in forecasting the biological activity based on the sample set PDGFR (platelet derived growth factor receptor) Inhibitors have shown its effectiveness.