PREDICTION PROBLEM: A LITERATURE REVIEW

1. Tổng quan về bài toán dự đoán

Ngày nay, dữ liệu được xem là tài sản quý giá của nhân loại bởi dữ liệu chứa nhiều thông tin tìm ẩn rất giá trị. Câu hỏi đặt ra là " Làm thể nào để phân tích, khai thác dữ liệu nhằm hỗ trợ con người rút ra được những thông tin hữu ích, từ đó đưa ra những quyết định đúng đắn trong công việc?". Để trả lời câu hỏi đó, bài toán dự đoán (prediction) ra đời. Dự đoán là một quá trình phân tích dữ liệu, sau đó, xây dựng các mô hình mô tả dữ liệu và dự đoán xu hướng trong tương lai **[1]**. Nhiệm vụ chính của dự đoán là xây dựng những mô hình dự đoán từ dữ liệu lịch sử. Dựa vào mô hình đó, chúng ta có thể gán nhãn các đối tượng dữ liệu (phân lớp), tính toán giá trị gần đúng của một giá trị dữ liệu trong tương lai, cho biết mối quan hệ của các đối tượng,…

Có nhiều ứng dụng của bài toán dự đoán như xây dựng mô hình dự đoán rủi ro khi cho vay của ngân hàng, dự đoán khả năng mắc bệnh trong y học, dự đoán về giá cả, tài chính, chứng khoán hay dự đoán nghề nghiệp cho học sinh, sinh viên,…

Bài toán dự đoán có hai dạng phổ biến: dự đoán tổng quát (prediction), dự đoán có yếu tố thời gian (forecasting). Đối với dạng bài toán dự đoán tổng quát, các nhà nghiên cứu chủ yếu sử dụng các phương pháp máy học **[1]**, thống kê **[2]**, mô hình tuyến tính **[15]**,… Các phương pháp này chủ yếu phân tích dữ liệu và xây dựng các mô hình phân lớp, mô hình tính toán giá trị tương lai,… Dạng bài toán dự đoán có tính thời gian được đề cập trong các công trình nghiên cứu **[3] [4] [8]**. Bài toán forecasting được giải quyết cũng dựa vào các phương pháp giống như prediction, tuy nhiên xem xét kỹ yếu tố thời gian. Chẳng hạn, kỹ thuật phân lớp theo thời gian (classification according to times-scales). Người ta chia những khoảng thời gian xác định để phân lớp như 1 phút đến 1 giờ, 1 giờ đến vài giờ, vài giờ đến một tuần,… Hoặc áp dụng thống kê, mạng nơ ron để tính toán giá trị dữ liệu ở một thời gian nhất đinh t, t+1, t+2,…

Phần tiếp theo sẽ trình bày cụ thể các phương pháp, kỹ thuật được sử dụng phổ biến để giải quyết bài toán dự đoán như thống kê, hồi quy, máy học,…

2. Các hướng tiếp cận cho bài toán dự đoán

2.1. Hướng tiếp cận dựa trên thống kê (statistical model)

Dự đoán dựa trên thống kê là một hướng tiếp cận truyền thống gắn liền với các định lý xác suất, mà tiêu biểu là định lý Bayes. Để áp dụng xác suất thống kê vào bài toán dự đoán, đối tượng dữ liệu phải có các thuộc tính và giả sử rằng các thuộc tính độc lập với nhau và có độ quan trọng như nhau. Chẳng hạn, dự báo thời tiết xem xét vấn đề "**có mưa hay không mưa**" cần có các thông tin về quang cảnh (outlook), nhiệt độ (temperature), độ ẩm (humidity), gió (windy). Từ những thông tin này, chúng ta sẽ dựa vào mô hình xác suất để tính xem với outlook, temperature, humidity, windy cụ thể thì xác suất có mưa sẽ như thế nào. Công thức tính xác suất dựa vào định lý Bayes:

P(H|E) = P(E|H) * P(H) / P(E)

Trong đó, **P(H|E)** *là xác suất xảy ra sự kiện H khi sự kiện E đã xảy ra*

P(E|H) *là xác suất xảy ra sự kiện E khi sự kiện H đã xảy ra*

P(H) *là xác suất xảy ra sự kiện H*

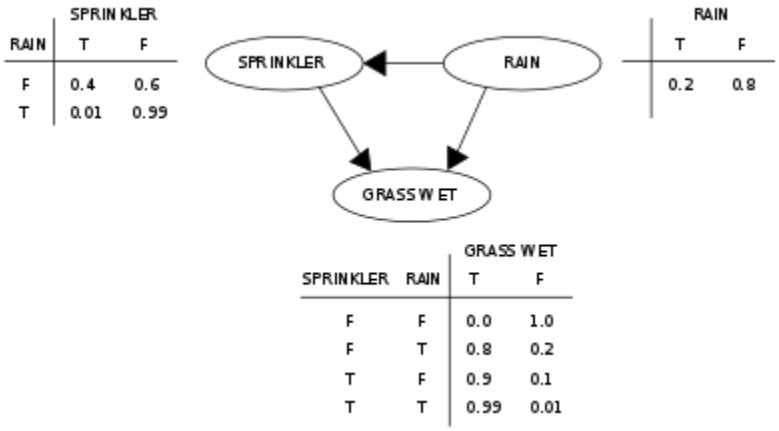
P(E) *là xác suất xảy ra sự kiện E*

Ưu điểm của hướng tiếp cận này là dễ dàng áp dụng, có thể đưa ra dự đoán nhanh chóng, áp dụng trên dữ liệu số (phải chuẩn hóa và rời rạc hóa) hay rời rạc đều được. Tuy nhiên, hướng tiếp cận này cũng có nhiều khuyết điểm:

- Độ chính xác dự đoán thường không cao đối với các bài toán phức tạp đòi hỏi độ chính xác dự đoán cao.
- Không áp dụng được trên dữ liệu liên tục, dữ liệu số.
- Do dựa trên xác suất nên có thể có một số trường hợp xác suất bằng 0 hoặc quá nhỏ (~0) gây khó khăn khi cho kết quả dự đoán.
- Lựa chọn các thuộc tính để làm các sự kiện tính xác suất như thế nào là phù hợp (chủ yếu dựa trên ý kiến chuyên gia).
- Nếu có quá nhiều thuộc tính sẽ gây ra vấn đề cho tính toán xác suất.

- Một số trường hợp ước tính trước xác suất của một sự kiện để suy ra xác suất xảy ra của sự kiện khác là rất khó.

Một cải tiến của hướng tiếp cận thống kê là sử dụng Bayesian Network. Mạng Bayes kết hợp xác suất với các mối quan hệ giữa các thuộc tính (thuộc tính này suy ra thuộc tính kia). Cấu trúc mạng Bayes gồm các đỉnh là các thuộc tính, các cạnh thể hiện mối liên hệ giữa các thuộc tính và trên cạnh có trọng số là xác suất xảy ra sự kiện H khi có sự kiện E (các sự kiện là các thuộc tính).



Hình 1. Minh họa Bayesian Network[1]

2.2. Hướng tiếp cận dựa trên mô hình hồi quy

Mô hình hồi quy là một phương pháp phân tích quan hệ giữa biến phụ thuộc Y với một hay nhiều biến độc lập X, từ đó xây dựng một mô hình (thường là một hàm số) để phân lớp các đối tượng. Các tham số của mô hình (hàm số) được ước lượng từ dữ liệu. Hướng tiếp cận dựa trên mô hình hồi quy sử dụng cho các đối tượng dữ liệu có thuộc tính kiểu số, kỹ thuật dự đoán phổ biến là hồi quy tuyến tính (linear regression). Khi dự đoán, người ta tính giá trị lớp đối tượng được dự đoán là kết hợp tuyến tính của các giá trị: C = w0*a0 + w1*a1 + w2*a2 + ... + wk*ak với w_i là hệ số nhân của các thuộc tính của đối tượng dữ liệu. Giá trị w_i được gán ngẫu nhiên trước, sau đó dựa vào dữ liệu, người ta sẽ

[1] https://en.wikipedia.org/wiki/Bayesian_network

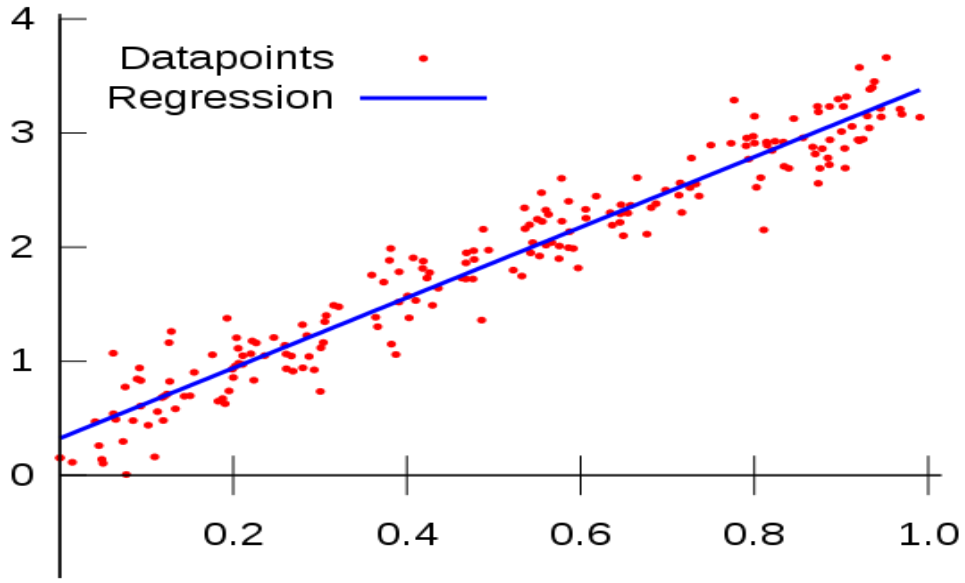
tìm độ lỗi toàn phương trung bình (squared error) khi dự đoán. Độ lỗi toàn phương trung bình bằng tổng lỗi của tất cả đối tượng dữ liệu chia cho số đối tượng dữ liệu dự đoán:

$$\text{MSE} = \frac{1}{n}\sum_{i=0}^{n}(Y' - Y)^2$$

Trong đó, Y' là giá trị dự đoán của mô hình

Y là giá trị thật từ dữ liệu

Sau khi tính được độ lỗi trung bình, người ta sẽ điều chỉnh các trọng số w_i sao cho độ lỗi toàn phương trung bình dưới một giá trị cho phép do người dùng qui định, thường độ lỗi MSE phải nhỏ hơn 0.05, 0.01,… là chấp nhận được. Do đó, có thể nói mô hình cần xây dựng bản chất là tìm được các trọng số w_i phù hợp với dữ liệu training.



Hình 2. Minh họa mô hình hồi quy tuyến tính

Tóm lại, các bước dự đoán áp dụng hồi quy tuyến tính:

- **Bước 1**: Xác định giá trị các thuộc tính của đối tượng dữ liệu.
- **Bước 2**: Khởi tạo giá trị trọng số w_i cho các thuộc tính của đối tượng dữ liệu để xây dựng mô hình hồi quy đầu tiên.
- **Bước 3**: Tính giá trị dự đoán dựa vào mô hình khởi tạo cho tập training.

- **Bước 4**: Điều chỉnh trọng số w_i sao cho giảm độ lỗi MSE.

Một phương pháp dự đoán cái tiến của hồi quy tuyến tính được trình bày ở công trình **[15]**. Mô hình ở đây được phát triển là mô hình MLS (multiple linear regression).

2.3. Hướng tiếp cận dựa trên mô hình phân lớp

Mô hình phân lớp là một phương pháp được sử dụng phổ biến trong data mining. Bài toán phân lớp là quá trình phân lớp một đối tượng dữ liệu vào một hay nhiều lớp đã cho trước nhờ một mô hình phân lớp (classification model). Mô hình này được xây dựng dựa trên một tập dữ liệu được xây dựng trước đó có gán nhãn (tập huấn luyện). Quá trình phân lớp là quá trình gán nhãn cho đối tượng dữ liệu. Như vậy, nhiệm vụ của bài toán phân lớp là cần tìm một mô hình phần lớp để khi có dữ liệu mới thì có thể xác định được dữ liệu đó thuộc vào phân lớp nào. Có nhiều bài toán phân lớp dữ liệu như phân lớp nhị phân (binary), phân lớp đa lớp (multiclass), phân lớp đa trị.

Việc xây dựng mô hình phân lớp là tìm ra hàm f(x) và thông qua hàm f tìm được để gán nhãn cho dữ liệu mới. Bước này thường được gọi là học hay training. Thông thường để xây dựng mô hình phân lớp cho bài toán này chúng ta sử dụng các thuật toán học giám sát (supervised learning) như KNN, Neural Network, SVM, Decision Tree.

Mô hình phân lớp thể hiện qua hàm số $y = f(x)$

Trong đó: x là các feature hay input đầu vào của dữ liệu

y là nhãn lớp hay output đầu ra.

Các bước thực hiện dự đoán với hướng tiếp cận dựa trên mô hình phân lớp:

- **Bước 1**: Chuẩn bị tập dữ liệu huấn luyện (dataset) và rút trích đặc trưng (feature extraction).
- **Bước 2**: Xây dựng mô hình phân lớp (classification model) bằng các thuật toán phân lớp.
- **Bước 3**: Kiểm tra dữ liệu với mô hình (make prediction).
- **Bước 4**: Đánh giá mô hình phân lớp và chọn ra mô hình tốt nhất.

2.3.1. Thuật toán kNN

Thuật toán kNN sử dụng ma trận tương tự để so sánh các thực thể (đối tượng dữ liệu) trong trong dữ liệu huấn luyện. Mỗi đối tượng dữ liệu sẽ có n đặc trưng. Để dự đoán nhãn cho một đối tượng chưa biết thì kNN sẽ chọn k đối tượng gần nhất với thực thể chưa biết và gán nhãn cho chúng. Để tìm các đối tượng gần nhất với thực thể chưa biết, kNN dùng độ đo khoảng cách Euclid. Có thể nói kNN là một thuật toán máy học "lười học (lazy learning)" bởi vì kNN không xây dựng trước mô hình từ dữ liệu huấn luyện mà chỉ tính toán khoảng cách giữa các đối tượng khi cần và đưa ra dự đoán.

Sau khi tìm ra k đối tượng gần nhất, kNN sử dụng kỹ thuật bỏ phiếu (xem xét trong k đối tượng gần nhất được lựa chọn, lớp nào có nhiều đối tượng nhất) để xem đối tượng dữ liệu chưa biết sẽ thuộc lớp nào. Một trong những lĩnh vực dự đoán sử dụng kNN là bài toán dự đoán giá cổ phiếu. Các bước dự đoán giá cổ phiếu sử dụng kNN **[20]** gồm:

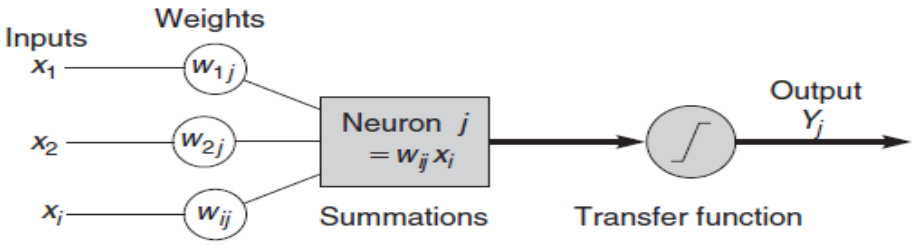
- **Bước 1**: Xác định số k hàng xóm gần nhất.
- **Bước 2**: Tính khoảng cách giữa các đối tượng dữ liệu trong tập huấn luyện và đối tượng cần phân lớp.
- **Bước 3**: Sắp xếp các đối tượng huấn luyện theo kết quả giảm dần khoảng cách với đối tượng cần phân lớp.
- **Bước 4**: Sử dụng kỹ thuật bỏ phiếu để gán nhãn cho đối tượng cần phân lớp.

2.3.2. Mạng neuron nhân tạo (Artificial Neural Network)

Mạng Neuron nhân tạo (Artificial Neural Network - ANN) là mô hình xử lý thông tin được mô phỏng dựa trên hoạt động của hệ thống thần kinh của sinh vật, bao gồm số lượng lớn các neuron được gắn kết để xử lý thông tin. ANN giống như bộ não con người, được học bởi kinh nghiệm (thông qua huấn luyện), có khả năng lưu giữ những kinh nghiệm hiểu biết (tri thức) và sử dụng những tri thức đó trong việc dự đoán các dữ liệu chưa biết (unseen data).

Kiến trúc chung của một ANN gồm 3 thành phần đó là Input Layer, Hidden Layer và Output Layer. Mỗi Input tương ứng với 1 thuộc tính (attribute) của dữ liệu. Ví dụ, trong

ứng dụng của ngân hàng xem xét có chấp nhận cho khách hàng vay tiền hay không thì mỗi Input là một thuộc tính của khách hàng như thu nhập, nghề nghiệp, tuổi, số con,… Output là kết quả mà mô hình ANN tính toán được, ví dụ, với bài toán xem xét chấp nhận cho khách hàng vay tiền hay không thì output là yes (cho vay) hoặc no (không cho vay).



Hình 3. Kiến trúc ANN

Connection Weights (trọng số liên kết) là thành phần rất quan trọng của một ANN, nó thể hiện mức độ quan trọng (trọng số) của dữ liệu đầu vào đối với quá trình xử lý thông tin (quá trình chuyển đổi dữ liệu từ layer này sang layer khác). Quá trình học của ANN thực ra là quá trình điều chỉnh các trọng số (Weight) của các input data để có được kết quả mong muốn (giảm sai lệch kết quả khi dự đoán). ANN dùng kỹ thuật lan truyền ngược (back-propagation network) để điều chỉnh trọng số giữa các neural và việc này tốn khá nhiều chi phí tính toán.

Summation Function (hàm tổng) có nhiệm vụ tính tổng trọng số của tất cả các input được đưa vào mỗi Neuron. Hàm tổng của một Neuron đối với n input được tính theo công thức sau:

$$Y=\sum_{i=1}^{n} Xi.Wi$$

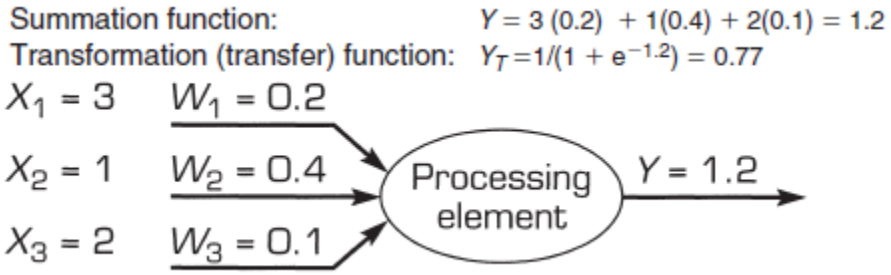
Transformation Function (hàm chuyển đổi) thể hiện mối quan hệ giữa các hidden neuron với kết quả output. Việc lựa chọn Transfer Function có tác động lớn đến kết quả của ANN. Hàm chuyển đổi phi tuyến được sử dụng phổ biến trong ANN là sigmoid (logical activation) function.

$$Y_T = 1/(1 + e^{-Y})$$

Trong đó : Y_T: Hàm chuyển đổi

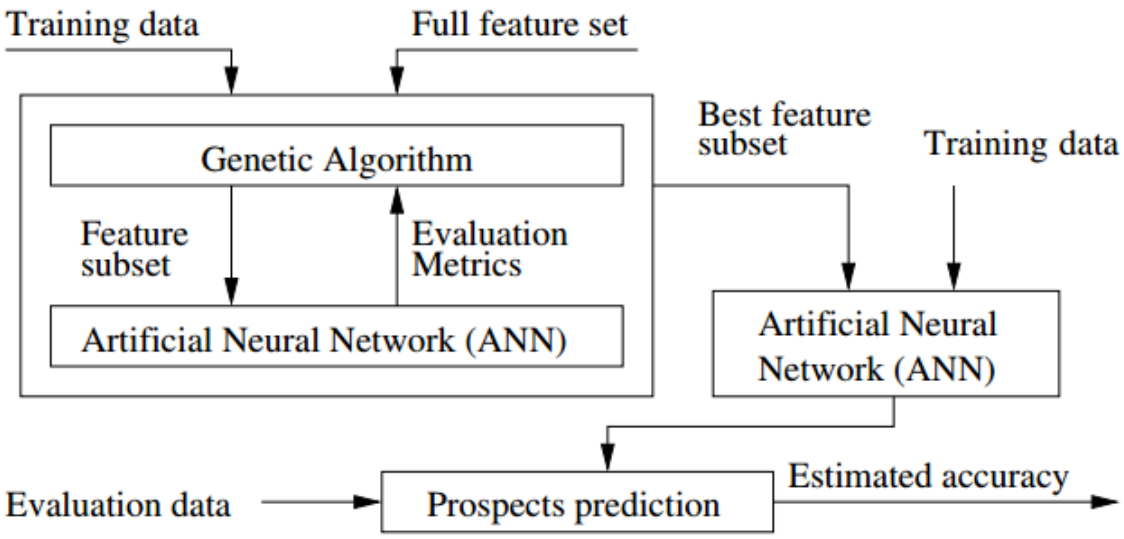
Y: Hàm tổng

Kết quả của Sigmoid Function thuộc khoảng [0,1] nên còn gọi là hàm chuẩn hóa (Normalized Function).



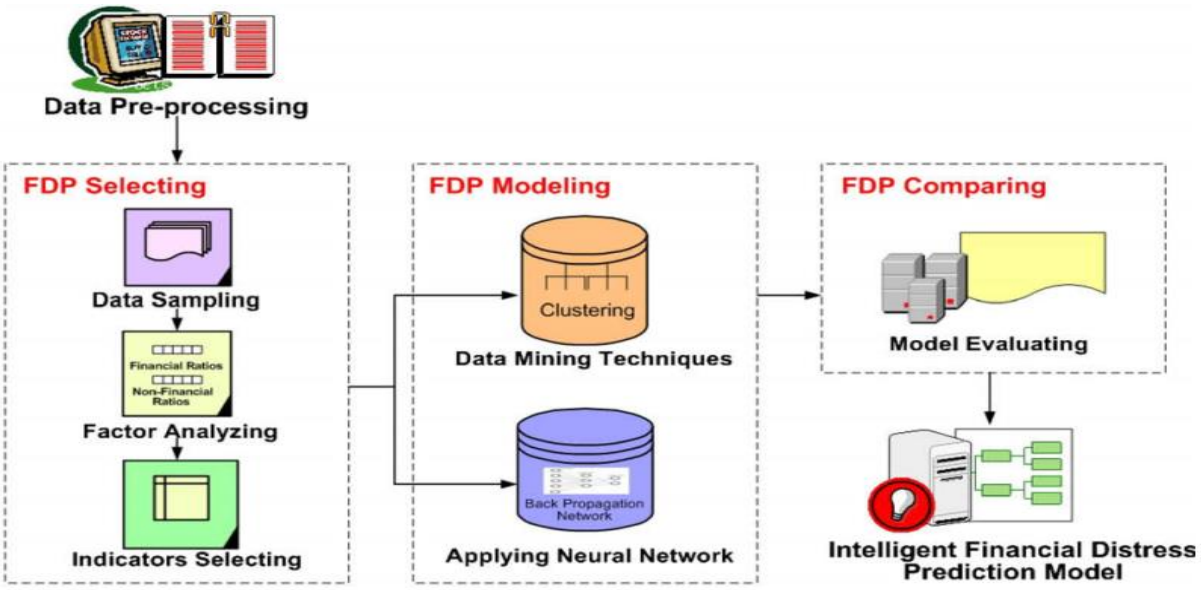
Hình 4. Minh họa hàm chuyển đổi trong ANN

Trong bài toán dự đoán, ANN được sử dụng trong nhiều lĩnh vực. ANN thường được kết hợp với các kỹ thuật khác để xây dựng hệ thống dự đoán hoàn chỉnh. Lĩnh vực quản lý thị trường, người ta cần nắm bắt nhu cầu, thị hiếu của người dùng từ dữ liệu mua hàng trước đó. Nhóm tác giả YongSeog Kim **[10]** đã kết hợp thuật toán di truyền (GA) và ANN để xây dựng hệ thống dự đoán nhu cầu đó. Mô hình gồm 2 thành phần cơ bản: lựa chọn tập đặc trưng và huấn luyện dữ liệu. Để lựa chọn đặc trưng, nhóm tác giả kết hợp GA và ANN. Sau đó, sử dụng ANN để huấn luyện tập dữ liệu dựa trên tập đặc trưng được rút trích để tạo thành mô hình dự đoán. Cuối cùng, thành phần dự đoán và đánh giá thuật toán được hình thành.



Hình 5. Cấu trúc mô hình GA/ANN

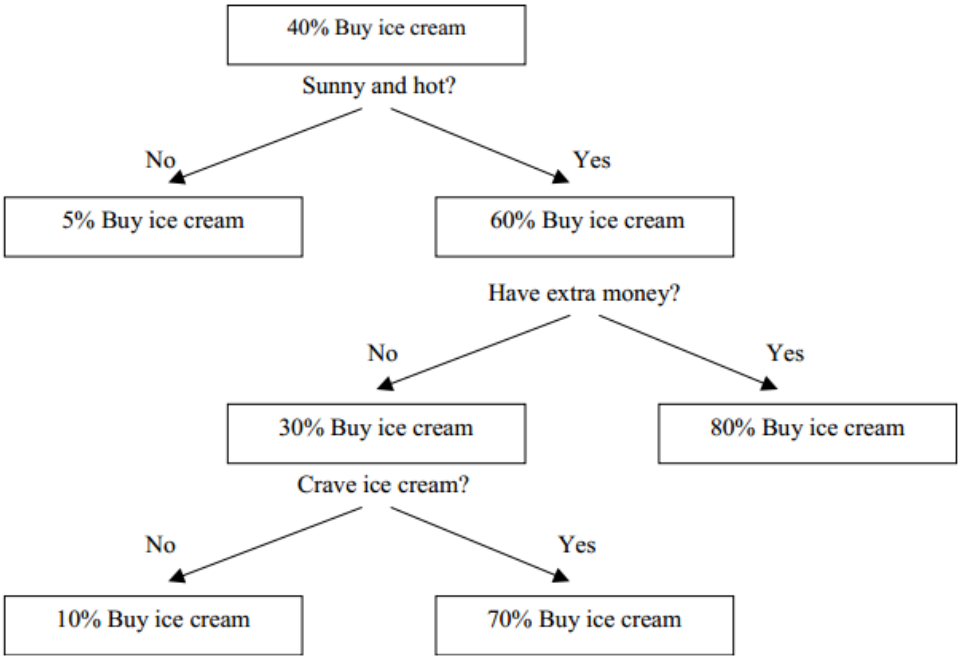
Wei-Sen Chen et al **[7]** kết hợp ANN và data mining để xây dựng mô hình dự đoán suy thoái tài chính ở Đài Loan, gọi là mô hình FDP. Mô hình này gồm 3 giai đoạn. Giai đoạn 1 là lựa chọn đặc trưng, giai đoạn 2 là xây dựng mô hình dự đoán FDP kết hợp cả ANN và kỹ thuật gom cụm trong data mining, giai đoạn 3 là bước đánh giá độ chính xác của mô hình, bản chất là so sánh độ chính xác dự đoán của ANN và kỹ thuật gom cụm.



Hình 6. Phương pháp xây dựng mô hình FDP

2.3.3. Cây quyết định

Cây quyết định là một kỹ thuật được sử dụng phổ biến trong data mining. Kỹ thuật này xây dựng các luật để chia các đối tượng dữ liệu thành từng nhóm riêng biệt. Nó là một thuật toán tiêu biểu cho mô hình phân lớp. Luật đầu tiên chia dữ liệu ra một vài phần nhỏ, sau đó một luật khác lại chia dữ liệu thành các phần nhỏ hơn, cứ như thế dữ liệu được chia thành những nhóm riêng biệt.



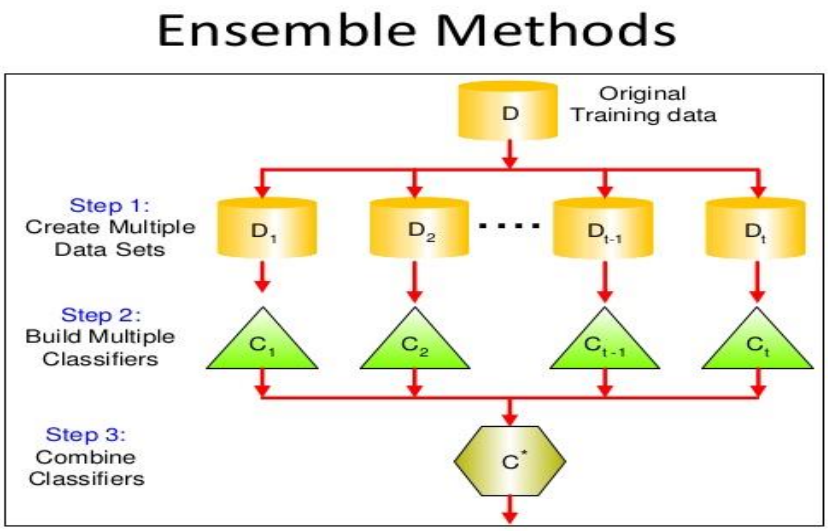
Hình 7. Minh họa cây quyết định **[12]**

Cây quyết định là một kỹ thuật dễ sử dụng để xây dựng mô hình dự đoán trong các lĩnh vực. Tuy nhiên, nó cũng có một số vấn đề cần giải quyết để nâng cao độ chính xác dự đoán:

- Lựa chọn các thuộc tính để xây dựng cây quyết định. Trường hợp có quá nhiều thuộc tính thì nút trên cây sẽ nhiều và mô hình dự đoán sẽ có nhiều luật phân lớp, dễ dẫn đến trường hợp overfitting dữ liệu, làm độ chính xác thuật toán giảm xuống thấp.
- Một số trường hợp dữ liệu trong tập huấn luyện bị thiếu, phải tìm cách điền vào các giá trị thiếu này để tiến hành huấn luyện mô hình và xây dựng cây quyết

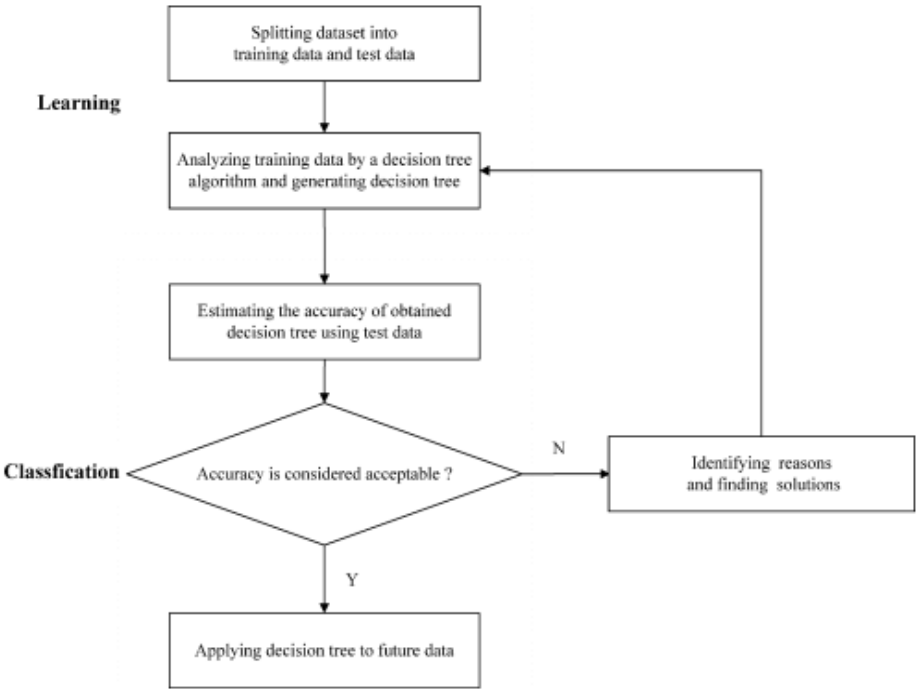
định. Vấn đề đặt ra là "Làm thế nào điền các giá trị dữ liệu thiếu đó mà vẫn đảm bảo không làm sai lệch mô hình huấn luyện?".

Các mô hình dự đoán hiện nay thường không chỉ sử dụng kỹ thuật cây quyết định mà còn có xu hướng kết hợp nhiều thuật toán dự đoán với nhau, người ta gọi phương pháp này là **ensembles**. Phương pháp này sẽ tiến hành kết hợp nhiều mô hình huấn luyện với nhau để tận dụng ưu điểm của các mô hình huấn luyện được tạo bởi các thuật toán khác nhau. Đối với một đối tượng dữ liệu cần dự đoán, đối tượng này sẽ được dự đoán bởi các mô hình huấn luyện khác nhau rồi tiến hành bỏ phiếu và chọn ra nhãn cho đối tượng dữ liệu này.



Hình 8. Minh họa Ensemble Methods

Năm 2010, Zhun Yu et al **[13]** sử dụng cây quyết định để xây dựng mô hình dự đoán nhu cầu năng lượng và đặt kết quả khá cao. Kết quả dự đoán cho thấy độ chính xác khá cao, độ chính xác 93% cho training data và 92% cho testing data.



Hình 9. Mô hình dự đoán nhu cầu năng lượng sử dụng cây quyết định **[13]**

2.4. Kỹ thuật gom cụm

K-Means là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật gom cụm. Tư tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng (objects) đã cho vào K cụm (K là số các cụm được xác đinh trước, K nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid) là nhỏ nhất. Các bước của thuật toán K-Means như sau:

- **Bước 1:** Xác định số k cụm cần được phân ra.
- **Bước 2:** Khởi tạo k cụm, thường chọn ngẫu nhiên k đối tượng dữ liệu làm phần tử trong mỗi cụm.
- **Bước 3:** Tính khoảng cách các đối tượng dữ liệu trong training data với các đối tượng trong mỗi cụm khởi tạo.
- **Bước 4:** Gán các đối tượng dữ liệu trong training data vào nhóm có khoảng cách gần nhất.

- **Bước 5:** Gán lại đối tượng trọng tâm của mỗi nhóm, tính lại và gán các đối tượng dữ liệu vào nhóm gần nhất với nó.

Thuật toán K-Means được sử dụng trong một số lĩnh vực như dự đoán khả năng học tập của sinh viên. Oyelade, O. J et al **[19]** đã sử dụng K-Means để dự đoán sinh viên nào học giỏi, học kém trong lớp.

2.5. Các kỹ thuật hiện đại (state of art)

Một số kỹ thuật mới được sử dụng trong các bài toán dự đoán như LSVM trong dự đoán HIV **[16]** với dữ liệu là các thông tin về một octamer (chuỗi amio axit). Bài toán này dự đoán liệu một octamer có bị cắt hay không ở vị trị thứ tư và thứ năm. Các kỹ thuật dự đoán gần đây phát triển mạnh trong lĩnh vực dự đoán liên kết (link prediction) trên mạng xã hội. WANG Peng **[17]** đã tổng kết lại các kỹ thuật dự đoán liên kết hiện đại nhất và chỉ ra một số thách thức và xu hướng tương lai. Ngoài ra, dự đoán chứng khoán cũng được J. G. Agrawal **[18]** giới thiệu các kỹ thuật hiện đại nhất hiện nay, chỉ ra các phân tích về các kỹ thuật thường dùng.

3. Các phương pháp dự đoán có tính thời gian

Như đã đề cập, dự đoán mang tính thời gian là một lĩnh vực dự đoán quan trọng. Các lĩnh vực áp dụng phương pháp dự đoán này như dự báo thời tiết **[8]**, dựa báo giá cả,…Các kỹ thuật được sử dụng cũng tương tự như bài toán dự đoán thông thường **[4]**, tuy nhiên có xem xét kỹ yếu tố thời gian. Một số phương pháp cho những bài toán dạng này có thể kể đến như:

- Phương pháp thống kê: xây dựng mô hình thống kê từ dữ liệu lịch sử. Về cơ bản, thống kê chỉ tốt khi dự đoán trong một khoảng thời gian ngắn. Vấn đề lớn nhất là độ lỗi sẽ càng lớn khi thời gian dự đoán lớn **[8]**.
- Áp dụng trí thông minh nhân tạo: có nhiều phương pháp mới được áp dụng như mạng neural nhân tạo (ANN), hệ thống suy diễn mờ (adaptive neuro-fuzzy inference system), máy hỗ trợ vector (SVM),… Các phương pháp này chủ yếu xây dựng các mô hình học, tự thay đổi các trọng số để xây dựng mô hình dự đoán chính xác nhất.

- Các phương pháp lai: kết hợp các phương pháp đã có nhằm nâng cao độ chính xác dự đoán chẳng hạn kết hợp ANN và phương pháp thống kê.

TÀI LIỆU THAM KHẢO

[1] Jiawei Han Jian Pei Micheline Kamber. *Data Mining, Southeast Asia Edition*. Morgan Kaufmann, 2006.

[2] Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2013.

[3] Tim Rey, Arthur Kordon, Chip Wells. *Applied Data Mining for Forecasting Using SAS*. SAS Institute, 2012.

[4] Timothy D. Rey, Chip Wells, Justin Kauhl (2013), "*Using Data Mining in Forecasting Problems*". SAS Global Forum in Data Mining and Text Analytics 2013, p. 085–2013.

[5] Abeer Badr El Din Ahmed, Ibrahim Sayed Elaraby (2014), "*Data Mining: A prediction for Student's Performance Using Classification Method"*. World Journal of Computer Application and Technology Vol. 2(2), pp. 43 – 47.

[6] Surjeet Kumar Yadav, Saurabh Pal (2012), "*Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification*". World of Computer Science and Information Technology Journal (WCSIT), Vol. 2, No. 2, pp. 51-56.

[7] Wei-Sen Chen, Yin-Kuan Du (2009), "*Using neural networks and data mining techniques for the financial distress prediction model*". Expert Systems with Applications, Volume 36, Issue 2, Part 2, March 2009, Pages 4075-4086.

[8] Wen-Yeau Chang (2014), "*A Literature Review of Wind Forecasting Methods*". Journal of Power and Energy Engineering, pp. 161-168.

[9] P. Ranjeet Kumar, R. Ramesh, T.Venkat Narayana Rao, Shireesha Dara (2013), "*Software Quality Prediction A Review and Current Trends"*. International Journal Of Engineering And Computer Science, Volume 2 Issue 4 April, 2013 Page No. 1147-1155.

[10] YongSeog Kim, W. Nick Street (2004), "*An Intelligent System for Customer Targeting:A Data Mining Approach"*. Journal Decision Support Systems, Volume 37 Issue 2, May 2004 Pages 215 – 228.

[11] M.Akhil jabbar, Dr.Priti Chandra, Dr.B.L Deekshatulu (2012), "*Heart Disease Prediction System using Associative Classification and Genetic Algorithm*". International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT.

[12] Padraic G. Neville (1999), "*Decision Trees for Predictive Modeling"*. SAS Institute Inc. 4 August 1999.

[13] Zhun Yu, Fariborz Haghighat, Benjamin C.M. Fung, Hiroshi Yoshino (2010), "*A decision tree method for building energy demand modeling*". Energy and Buildings, pp. 1637–1646.

[14] Hai-xiang ZHAO, Frédéric MAGOULÈS (2012), "*Feature Selection for Predicting Building Energy Consumption Based on Statistical Learning Method*". Journal of Algorithms & Computational Technology Vol. 6 No. 1.

[15] Gregory T. Knofczynski and Daniel Mundfrom (2007), "*Sample Sizes When Using Multiple Linear Regression for Prediction*". Educational and Psychological Measurement, Volume 68 Number 3, pp. 431-442.

[16] Thorsteinn Rognvaldsson, Liwen You, Daniel Garwicz (2015). "*State of the art prediction of HIV-1 protease cleavage sites"*. Bioinformatics, pp. 1204–1210.

[17] WANG Peng, XU BaoWen, WU YuRong, ZHOU XiaoYu (2015), "*Link Prediction in Social Networks: the State-of-the-Art*". SCIENCE CHINA Information Sciences, January 2015, Vol. 58.

[18] J. G. Agrawal, V. S. Chourasia, A. K. Mittra (2013), "*State-of-the-Art in Stock Prediction Techniques*". International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 2, Issue 4.

[19] Oyelade, Oladipupo, Obagbuwa (2010), "*Application of k-Means Clustering algorithm for prediction of Students' Academic Performance*". International Journal of Computer Science and Information Security, Vol. 7, No. 1.

[20] Khalid Alkhatib, Hassan Najadat, Ismail Hmeidi, Mohammed K. Ali Shatnawi, "*Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm*". International Journal of Business, Humanities and Technology, Vol. 3 No. 3, 2013.