ỨNG DỤNG KỸ THUẬT GOM CỤM VÀ MẠNG NƠRON TRONG DỰ BÁO ĐIỂM TỐT NGHIỆP TRUNG HỌC PHỔ THÔNG TẠI TÂY NINH

APPLYING CLUSTERING TECHNIQUES AND NEURAL NETWORKS TO FORECAST THE HIGH SCHOOL'S GRADUATION SCORES IN TAY NINH PROVINCE

Đặng Trường Sơn,
ĐH Sư Phạm Kỹ Thuật TP.HCM.
Phạm Nguyễn Diễm Phú,
ĐH CNTT, ĐHQG-TP.HCM

TÓM TẮT

Mạng nơron nhân tạo đã được ứng dụng trong giải quyết nhiều bài toán thực tế. Mạng nơron truyền thẳng với thuật toán lan truyền ngược là mô hình phổ biến và thường được sử dụng trong giải quyết các bài toán dự báo. Bài báo đề xuất ứng dụng mô hình kết hợp mạng nơron truyền thẳng với thuật toán gom cụm K-means để xây dựng công cụ dự báo điểm thi tốt nghiệp Trung học phổ thông (THPT) tại tỉnh Tây Ninh.

ABSTRACT

Artificial neural networks have been applied in solving many practical problems. Neural network algorithms are widely accepted and often used in solving forecasting problems. This paper proposed a model that combines neural networks with K-means clustering algorithms to build a software application predicting high school graduation scores in Tay Ninh province.

I. ĐẶT VẤN ĐỀ

Tỷ lệ đậu tốt nghiệp Trung học phổ thông (THPT) hằng năm tại một tỉnh/ thành phố phần nào phản ánh chất lượng đào tạo tại địa phương đó. Vì vậy nếu được dự báo chính xác, các đơn vị giáo dục sẽ có những biện pháp cải tiến, điều chỉnh công tác dạy và học để cải thiện tỷ lệ tốt nghiệp. Thông thường giáo viên bộ môn thông qua việc đánh giá trình độ học sinh mình giảng dạy có thể dự báo kết quả thi tốt nghiệp của các em. Ban giám hiệu nhà trường tổng hợp dự báo của các giáo viên bộ môn để làm dự báo cho toàn trường. Sở Giáo dục & đào tạo hoàn toàn có thể dùng kết quả từng trường để làm dự báo chung cho cả tỉnh. Vấn đề hiện nay là giáo viên dạy quá nhiều lớp, số lượng học sinh lớn vì vậy khó có thể dự báo chính xác cho tất cả học sinh. Bên cạnh đó đánh giá của từng giáo viên sẽ mang nhiều tính chủ quan, tại một thời điểm cụ thể. Và cuối cùng là việc đánh giá như vậy khá thủ công, tốn nhiều công sức, thời gian.

Do đó, mục đích của bài báo này là giới thiệu một công cụ đã được xây dựng thử nghiệm để hỗ trợ việc dự báo điểm thi tốt nghiệp cho các nhà quản lý giáo dục tại Tây Ninh.

Bài toán dự báo điểm thi tốt nghiệp sử dụng dữ liệu điểm học và điểm thi các năm học trước ở trường để làm dữ liệu dự đoán điểm thi cho năm tới. Dữ liệu điểm cũ được sử dụng như là dữ liệu học để tìm ra những qui luật giữa điểm học và điểm thi. Kết quả của việc học đó là một bộ trọng số dùng để dự đoán điểm thi trong năm tới.

Chẳng hạn cần dự báo điểm thi tốt nghiệp môn Vật lý năm học 2008/09 của học sinh trường THPT Lý Thường Kiệt – Tây Ninh. Công cụ lấy dữ liệu học là điểm thi tốt nghiệp của các học sinh năm học 2007/08, 2006/07 và điểm trung bình môn từng học kỳ của các học sinh đó trong 3 năm lớp 10, 11, 12.

Ví dụ: Dữ liệu điểm học và điểm thi môn Vật lý của học sinh trường Lý Thường Kiệt 2 năm học 2007/08, 2006/07 như sau:

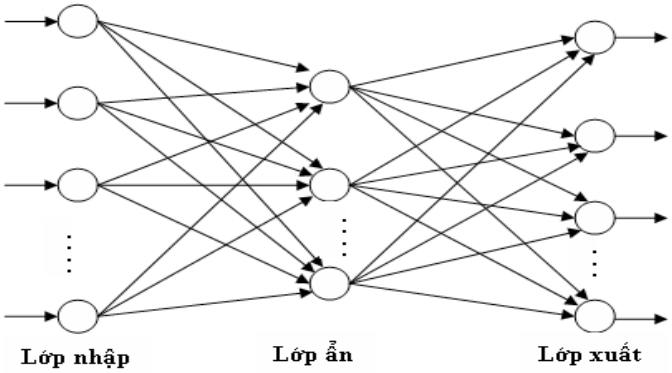
MaHS	TBM HK.I Lớp 10	TBMHK.II Lớp 10	TBM HK.I Lớp 11	TBMHK.II Lớp 11	TBM HK.I Lớp 12	TBMHK.II Lớp 12	Điểm thi TN
0001	5.2	5.5	5.7	3.5	6.1	5.2	5.5
0002	4.1	4.5	5.7	6.3	5.4	5.9	4.0
0003	5.0	5.1	6.3	6.1	6.1	6.2	4.5
0004	4.2	4.7	4.9	5.1	6.3	5.7	4.5
0005	6.1	4.5	3.8	5.8	5.7	5.1	6.5
0006	4.2	4.3	5.6	6.7	5.9	6.7	4.5
0007	6.7	6.9	7.2	6.7	6.9	6.2	6.5
....	.....	....	....	....	....	....	....
2000	5.1	4.8	4.7	4.3	5.0	5.1	5.5

Bảng 1: Bảng dữ liệu mẫu điểm môn Lý của học sinh trường Lý Thường Kiệt

Với dữ liệu điểm môn Vật lý của một học sinh trong năm học cần dự báo (2008/09) tương ứng với các cột của bảng trên, ví dụ: 6.7; 6.3; 6.4; 6.1; 7.6; 7.9, chương trình sẽ dự báo điểm thi tốt nghiệp môn Lý của học sinh này là bao nhiêu?

II. DỰ BÁO ĐIỂM THI TỐT NGHIỆP SỬ DỤNG MẠNG NƠRON

Mạng nơron [1- 3] là thuật ngữ nói đến một phương pháp giải quyết vấn đề - bài toán trên máy tính, mô phỏng theo hoạt động của các tế bào thần kinh trong não bộ. Mạng được tạo thành bởi sự nối kết giữa rất nhiều đơn vị thần kinh gọi là perceptron. Những đơn vị này có nhiệm vụ nhận các tín hiệu từ các đơn vị khác hoặc từ dữ liệu nhập; thông qua các mối liên kết, đơn vị sẽ tiến hành tổng hợp tín hiệu đến nó, xử lý và truyền các tín hiệu sang các đơn vị thần kinh khác hoặc đến đầu ra. Mạng nơron truyền thẳng có khả năng học rất tốt trên thuật toán lan truyền ngược.



Hình 1: Mạng nơron truyền thẳng 3 lớp

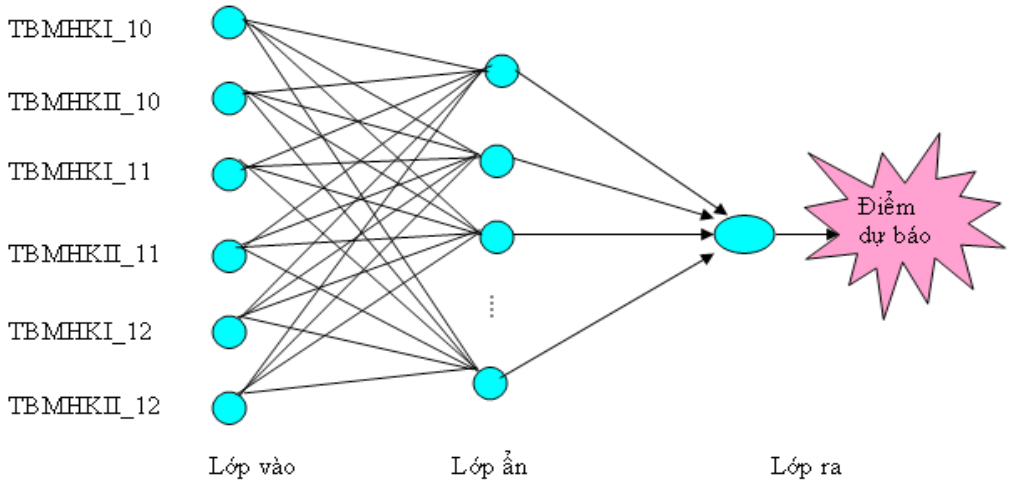
Hàm truyền thường được sử dụng trong quá trình luyện mạng là hàm logistic

$$g(x) = \frac{1}{1+e^{-x}}$$

Mạng nơron hoạt động ở hai trạng thái:

- Trạng thái ánh xạ, thông tin lan truyền từ lớp nhập qua lớp ẩn đến lớp xuất và mạng thực hiện ánh xạ để tính giá trị các biến phụ thuộc dựa vào các giá trị biến độc lập.
- Trạng thái học, thông tin lan truyền theo hai chiều nhiều lần để học các trọng số.

Mạng nơron dự báo điểm tốt nghiệp được thiết kế với 6 nút đầu vào và 1 nút đầu ra:



Hình 2: Mô hình mạng nơron dự báo điểm tốt nghiệp THPT

Các thông số kỹ thuật của mạng như sau:

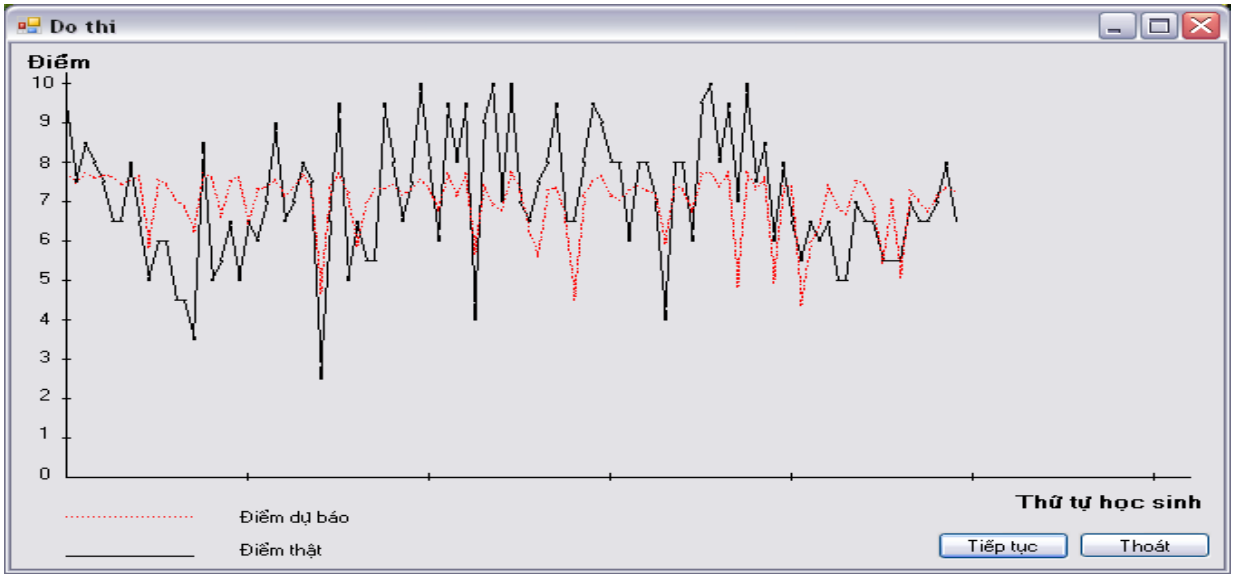
- Là mạng lan truyền ngược một lớp ẩn. Các hệ số học thay đổi theo hệ số học thích nghi
- Số nút nhập: 6 nút, tương ứng với 6 cột điểm đã nêu ở phần trên
- Số nút xuất: 1 nút, tương ứng với điểm dự đoán
- Số nút ẩn được quyết định bởi người luyện mạng
- Mạng dùng qui tắc delta, qui tắc học theo từng mẫu và qui tắc học thích nghi delta – bar – delta để học. Sau đó so sánh cách học nào là tối ưu hơn sẽ được chọn
- Bộ trọng số được khởi tạo ngẫu nhiên trong đoạn [0,1] có kiểm soát
- Hàm truyền: sử dụng hàm sigmoid (Logistic) $g(x) = \frac{1}{1+e^{-x}}$
- Mạng được huấn luyện bởi thuật toán lan truyền ngược (Algorithm BackPropagation)

Kết quả thử nghiệm trên tập dữ liệu điểm môn Lý của học sinh trường Lý Thường Kiệt tốt nghiệp năm học 2006/07, 2007/08:

- Tập D gồm tổng cộng 734 mẫu
- Tập D_{XD} = 2/3 D : 529 mẫu *(Tập D_{XD} là tập dữ liệu để huấn luyện mạng)*
- Tập D_{KC} = 1/3 D: 264 mẫu *(Tập D_{KC} là tập dữ liệu kiểm chứng)*
- Sử dụng 9 bộ trọng số thử nghiệm và các ngưỡng lỗi tương ứng
- Áp dụng qui tắc giảm dốc nhất (qui tắc Delta)

Bộ trọng số	1	2	3	4	5	6	7	8	9	Trung bình
Ngưỡng lỗi	0.02428	0.02506	0.02512	0.02432	0.02478	0.02507	0.02532	0.02425	0.02514	**0.02482**
Số mẫu đạt	89	77	76	77	88	90	94	88	92	**85.7**
Số mẫu trong tập D_{KC}										**264**
Tỷ lệ đạt *(Trung bình)*										**32.5%**

Bảng 2: Bảng thống kê kết quả dự báo theo qui tắc giảm dốc nhất

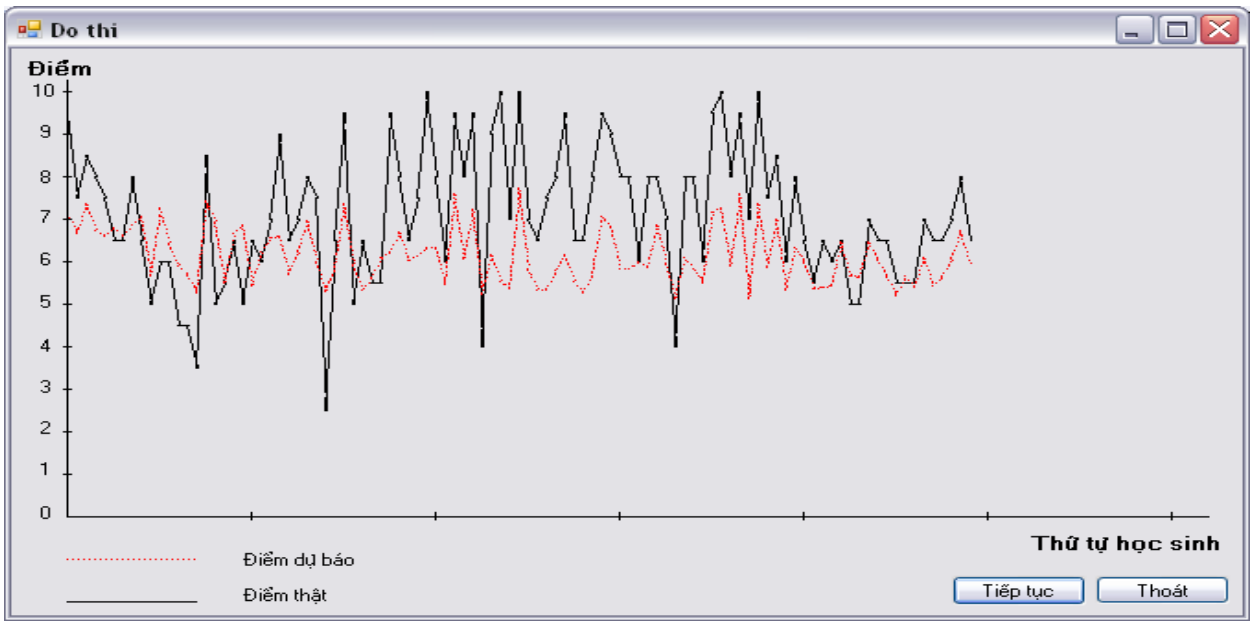


Hình 3: Đồ thị kết quả dự báo qui tắc giảm dốc nhất (bộ trọng số thứ 8)

Vẫn sử dụng tập dữ liệu trên để huấn luyện nhưng áp dụng qui tắc học từng mẫu.
Kết quả luyện mạng thu được như sau:

Bộ trọng số	1	2	3	4	5	6	7	8	9	**Trung bình**
Ngưỡng lỗi	0.03032	0.03050	0.03042	0.03121	0.03541	0.03214	0.03054	0.03781	0.03541	**0.03264**
Số mẫu đạt	117	122	116	132	145	155	165	103	140	**132.8**
Số mẫu trong tập D_{KC}										**264**
Tỷ lệ đạt *(Trung bình)*										**50.3%**

Bảng 3: Bảng thống kê kết quả dự đoán theo qui tắc học từng mẫu



Hình 3: Đồ thị kết quả dự báo theo qui tắc học từng mẫu (bộ trọng số thứ 2)

Qua đồ thị kết quả dự báo ở các cách học, dễ nhận thấy kết quả của đồ thị dự báo chỉ thể hiện được dạng của đồ thị điểm thật. Và từ các bảng thống kê cho thấy tỷ lệ số mẫu dự báo đạt chưa cao. Điều đó có thể giải thích bởi vì mạng được học với kết quả của các học sinh có quá nhiều khác biệt nên không thể tìm được qui luật chung cho tất cả. Mặc dù dữ liệu đầu vào trước khi học đã được tiền xử lý để loại bỏ đi các dữ liệu không mang tính qui luật nhưng việc gom các sinh viên theo từng nhóm dựa trên kết quả học để huấn luyện riêng rẽ sẽ có cơ may tìm được các bộ trọng số chính xác hơn.

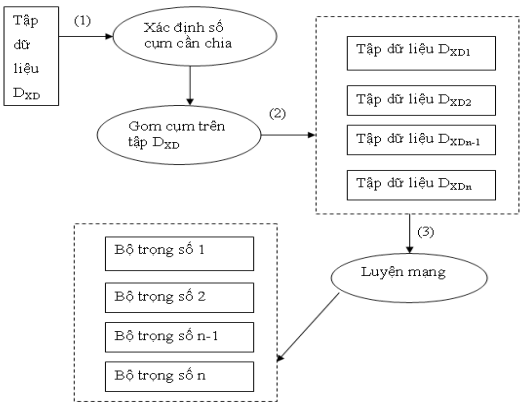
III. KẾT HỢP MẠNG NƠRON VỚI THUẬT TOÁN K-MEANS

Việc phân nhóm sẽ được thực hiện bằng thuật toán gom cụm K-means [4]. Quá trình giải bài toán bằng mô hình kết hợp mạng nơron truyền thẳng và thuật toán gom cụm K-means được chúng tôi thực hiện với các bước như sau:

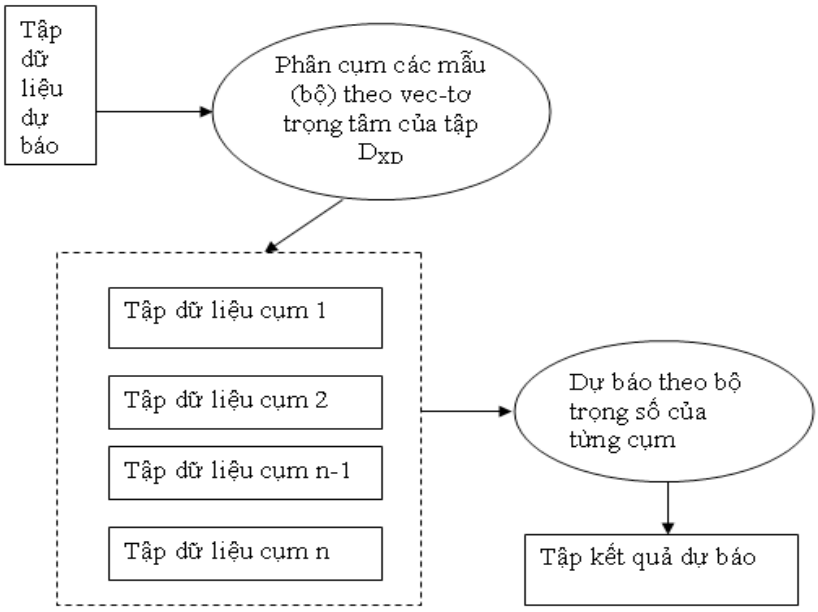
1) Xác định số cụm cần phân chia;

2) Gom cụm trên tập $D_{XD;}$

3) Phân lớp định lượng bằng mạng nơron truyền thẳng và thuật toán lan truyền ngược cho từng cụm.



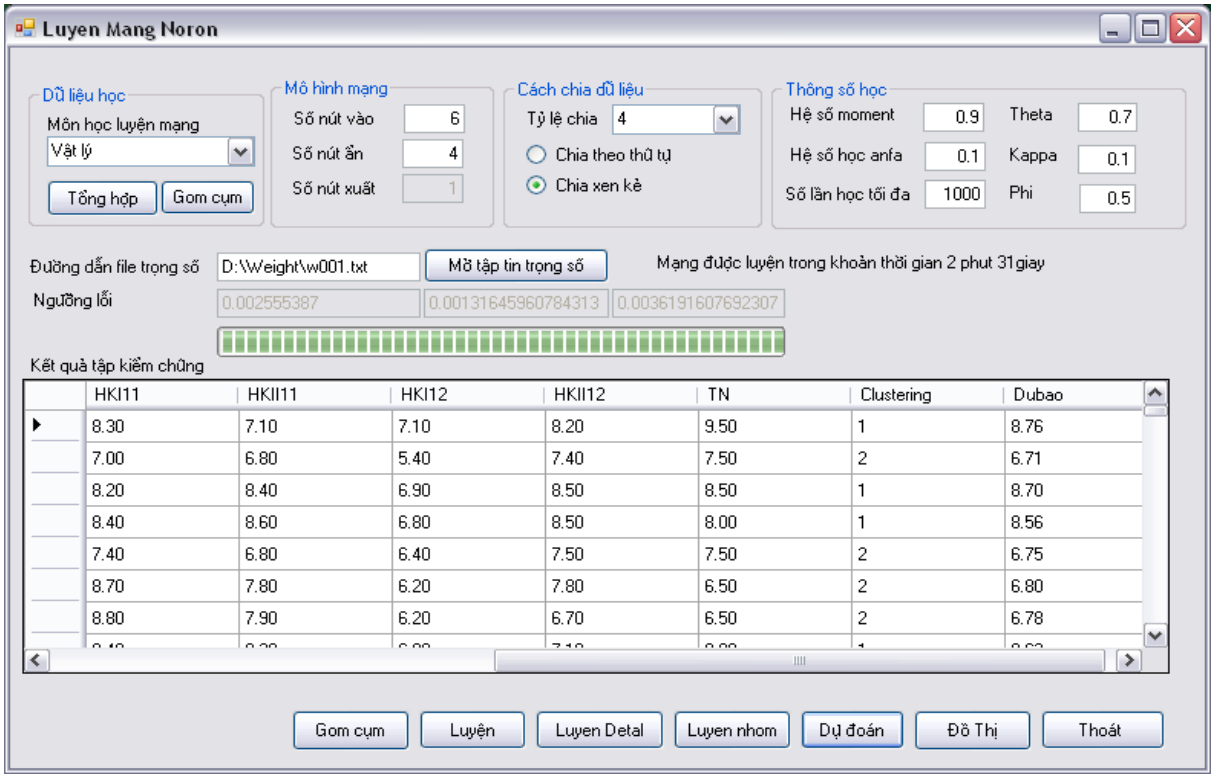
Hình 4: Mô hình các bước kết hợp gom cụm và mạng nơron



Hình 5: Mô hình các bước của quá trình dự báo

Như vậy dữ liệu trước khi học sẽ được gom cụm và sau đó sẽ được học riêng rẽ để cho kết quả là các bộ trọng số của riêng từng cụm. Dữ liệu dự báo cũng sẽ được phân vào các cụm khác nhau với các bộ trong số tương ứng để tính ra kết quả dự báo.

Giao diện người dùng của quá trình luyện mạng như sau:



Hình 6: Giao diện người dùng của chương trình

- Nhấp nút "**Tổng hợp"** để tổng hợp dữ liệu các năm học trước, sau đó nhấn nút "**Gom cụm"** để gom cụm trên tập dữ liệu vừa tổng hợp

- Chọn mô hình mạng, tỷ lệ chia tập dữ liệu xây dựng và tập dữ liệu kiểm chứng cũng như cách chia dữ liệu *(xen kẽ hay có thứ tự)*. Nhập các thông số học, đường dẫn lưu tập tin bộ trọng số. Cuối cùng nhấn nút **"Luyện"** *(Luyện Delta, Luyện nhóm)* tùy cách học. Kết thúc quá trình luyện mạng, chương trình cung cấp ngưỡng lỗi trong quá trình luyện mạng cũng như thời gian luyện mạng.

- Nhấp nút **"Dự đoán"** để dự báo kết quả trên tập kiểm chứng và trả kết quả trên lưới.

- Nhấp nút **"Đồ thị"** chương trình vẽ đồ thị kết quả dự báo và kết quả thật**.**

IV. PHÂN TÍCH KẾT QUẢ

Vẫn sử dụng tập dữ liệu trên để huấn luyện.

- Số cụm đề xuất: 3. Thực hiện gom cụm trên tập dữ liệu D_{XD} cho kết quả như sau:

  Cụm 1: 131 mẫu, cụm 2: 179 mẫu, cụm 3: 219 mẫu

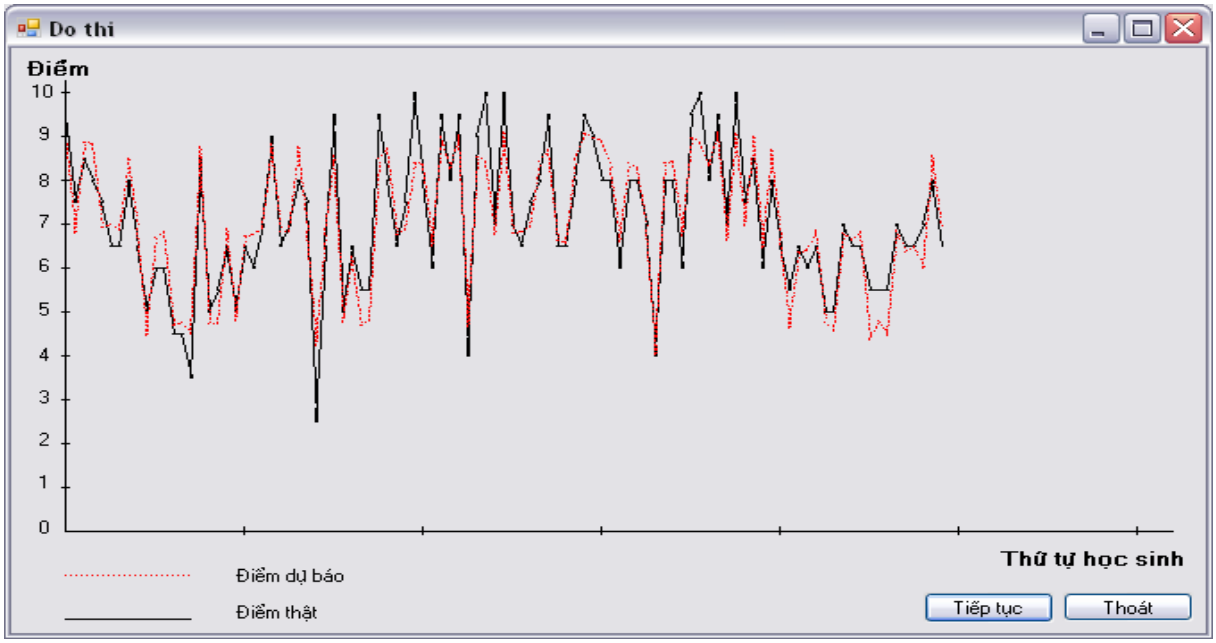
- Tập D_{KC} = 1/3 D: 264 mẫu

- Số bộ trọng số thử nghiệm: 7

Kết quả luyện mạng thu được kết quả như sau:

Bộ trọng số	Ngưỡng lỗi			Số mẫu đạt tập D_{KC}			Tổng số mẫu đạt
	Cụm 1	Cụm 2	Cụm 3	Cụm 1	Cụm 2	Cụm 3	
1	0.00779	0.00268	0.00450	23	84	126	233
2	0.00992	0.00165	0.00453	23	84	126	233
3	0.00162	0.00150	0.00410	23	85	125	233
4	0.00239	0.00150	0.00500	23	84	111	218
5	0.00500	0.00150	0.00470	23	84	126	233
6	0.00194	0.00145	0.00401	23	85	112	220
7	0.00188	0.00133	0.00392	19	82	123	224
Trung bình	0.00436	0.0017	0.00440	22.5	84	121.3	**227.7**
Số mẫu trong tập D_{KC}							**264**
Tỷ lệ đạt *(Trung bình)*							**86.3%**

Bảng4: Bảng thống kê kết quả dự báo bằng mạng nơron có kết hợp gom cụm



Hình 7: Đồ thị kết quả dự báo bằng cách kết hợp gom cụm và mạng nơron (bộ trọng số thứ 7)

Kết quả thử nghiệm trên cho thấy mô hình đề xuất đã dự báo với tỷ lệ chính xác khá tốt là 86.3%. Trong đồ thị trên có thể thấy các kết quả dự báo đã khá sát với kết quả thực tế. Như vậy việc kết hợp mạng nơron với kỹ thuật gom cụm dữ liệu đã cho chúng ta kết quả tốt hơn nhiều so với không gom cụm. Tuy nhiên kết quả trên vẫn chưa thể có giá trị thực tiễn trong việc dự đoán điểm thi của một học sinh cụ thể mà giá trị thực tiễn của nó nhằm dự đoán kết quả chung, có tính thống kê của một tập thể học sinh.

V. KẾT LUẬN

Nhiều mô hình mạng nơron khác nhau đã được nghiên cứu và ứng dụng trong các lĩnh vực cuộc sống đa dạng. Mạng nơron truyền thẳng dùng thuật toán lan truyền ngược tỏ ra hiệu quả trong xây dựng các công cụ dự báo. Tuy vậy việc dùng mạng nơ ron truyền thẳng để huấn luyện trên tập mẫu có nhiều đặc điểm khác nhau thì kém hiệu quả hơn do trong quá trình huấn luyện thông tin dễ bị nhiễu, hội tụ kém. Vì vậy giải pháp chia nhỏ tập mẫu huấn luyện thành các cụm nhỏ hơn, có những đặc điểm tương tự nhau để huấn luyện thì có hiệu quả hơn nhiều. Kết quả kiểm tra giải pháp này trên cùng một tập dữ liệu thử nghiệm cho thấy kết quả được cải thiện đáng kể so với trường hợp không phân nhóm.

TÀI LIỆU THAM KHẢO

Carlos Ordonez, "Integrating K-Means Clustering with a Relational DBMS Using SQL." *IEEE Transactions on Knowledge and Data Engineering*, vol.18, no.2., pp.188-201, 2006.

James A. Freeman, David M. Skapura, *Neural Networks Algorithms Applications and Programming Techniques*. Addison-Wesley Publishing Company.

Kenvin Gurney, *An Introduction to Neural Networks*. Taylor & Francis, 2003.

Trần Ngọc Anh, "Dùng mạng nơron và thuật giải di truyền để phân lớp dữ liệu." Luận văn thạc sĩ Trường ĐH KHTN, 2004.