Nội dung

[CÁC CÔNG VIỆC ĐÃ THỰC HIỆN 3](#_Toc291487137)

[VỀ MÔ HÌNH CẢI TIẾN ANN 3](#_Toc291487138)

[Mục đích 3](#_Toc291487139)

[Mô tả cải tiến 3](#_Toc291487140)

[THỰC NGHIỆM 3](#_Toc291487141)

[Mục đích 3](#_Toc291487142)

[Mô tả dữ liệu 3](#_Toc291487143)

[Kết quả 3](#_Toc291487144)

[KẾT LUẬN 7](#_Toc291487145)

[CÔNG VIỆC TƯƠNG LAI 7](#_Toc291487146)

Báo cáo kết quả dự đoán xu hướng – 24/04/11

# CÁC CÔNG VIỆC ĐÃ THỰC HIỆN

* Cài đặt và thử nghiệm mô hình ANN, SVM, SVM kết hợp xác suất.
* Cải tiến SVM với việc kết hợp Kmeans: KSVMeans.
* Cải tiến ANN với việc kết hợp với cây quyết định.
* Đề xuất độ đo mới để đánh giá lỗi mô hình: Sign Error Rate.

# VỀ MÔ HÌNH CẢI TIẾN ANN

## Mục đích

* Loại bỏ bớt những mẫu dữ liệu học có khả năng gây nhiễu cho ANN và chọn lựa ra những mẫu học với tính phân loại cao dữ trên các mối quan hệ của các tham số đầu vào
* Giảm thiểu tình trạng quá khớp khi bộ học số lượng lớn

## Mô tả cải tiến

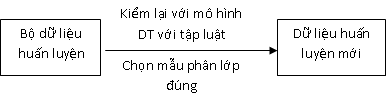
Mô hình cải tiến bao gồm 4 bước:

Bước 1: Từ bộ dữ liệu huấn luyện thực hiện quá trình học để xây dựng ra cây quyết định với tập luật rút ra



1. Xây dựng cây quyết định với bộ luật từ dữ liệu huấn luyện

Bước 2: Tạo bộ dữ liệu huấn luyện mới từ bộ dữ liệu cũ được phân lớp dúng với mô hình DT với bộ luật.



Dữ liệu kiểm thử

Kết quả

Kiểm thử trên mô hình ANN

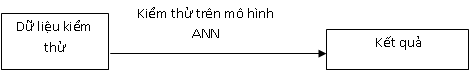
1. Chọn dữ liệu học mới với mô hình DT

Bước 3: Xây dựng mô hình ANN với bộ dữ liệu huấn luyện mới



1. Xây dựng mô hình ANN với dữ liệu huấn luyện mới

Bước 4: Thực hiện kiểm thử trên mô hình ANN với bộ dữ liệu kiểm thử và đánh giá kết quả



1. Kiểm thử với bộ dữ liệu kiểm thử trên mô hình ANN

# THỰC NGHIỆM

## Mục đích

* Đánh giá hiệu suất các phương pháp cải tiến so với các mô hình cơ sở.
* Quyết định kích thước bộ dữ liệu huấn luyện để đảm bảo độ chính xác dự đoán của các mô hình cao nhất.
* Ghi nhận thời gian huấn luyện giữa các phương pháp

## Mô tả dữ liệu

* Bao gồm các mã chứng khoán:
* Ở mỗi mã chứng khoán, ta đánh giá thông qua 3 bộ dữ liệu được tóm tắt như sau:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Khoảng thời gian | Tổng số mẫu | Số mẫu test |
| Bộ 1 | 04/01/2005-31/12/2008 | 997 | 200 |
| Bộ 2 | 05/01/2004-31/12/2008 | 1247 | 200 |
| Bộ 3 | 02/01/2003-31/12/2008 | 1494 | 200 |

## Kết quả

* Khảo sát số cụm k cho KSVMeans

Để khảo sát, ta chọn bộ 1 làm dữ liệu thử nghiệm cho phương pháp này với k mang giá trị 2, 3 và 4. Bảng sau thể hiện các kết quả thu được.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Mã** | **2-SVMeans** | | **3-SVMeans** | | **4-SVMeans** | |
| **Acc(%)** | **SERate(%)** | **Acc(%)** | **SERate(%)** | **Acc(%)** | **SERate(%)** |
| AGF | 71 | 0 | 60 | 1 | 58.5 | 6 |
| BT6 | 76.5 | 0 | 77.5 | 0 | 64 | 0 |
| GIL | 60.5 | 1 | 66.5 | 1.5 | 58.5 | 1 |
| GMD | 82 | 0 | 80.5 | 0 | 81 | 0 |
| HAP | 85.5 | 0 | 73 | 0 | 74 | 1 |
| KHA | 70 | 0 | 56 | 0 | 51 | 0 |
| LAF | 95.5 | 0 | 90 | 0 | 92 | 0 |
| REE | 85 | 0 | 73.5 | 0 | 74.5 | 0 |
| SAM | 75.5 | 0.5 | 70 | 1.5 | 65.5 | 8 |
| TRI | 90.5 | 0 | 79.5 | 0 | 88 | 0 |
| VNINDEX | 88.5 | 0 | 76.5 | 0 | 60 | 0 |
| **Trung bình** | ***80*** | ***0.14*** | ***73*** | ***0.36*** | ***69.72*** | ***1.45*** |

* Nhận xét: ta thấy K-SVMeans với k = 2 cho kết quả tốt nhất. Càng tăng k, độ chính xác càng giảm và độ lỗi càng tăng.
* Dự đoán 1 ngày trên 3 bộ dữ liệu

**Bộ 1**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Mã** | **BPNN** | | **DT-BPNN** | | **SVM** | | **SVM-Prob** | | **K-SVMeans** | |
| ***Acc*** | ***SERate*** | ***Acc*** | ***SERate*** | ***Acc*** | ***SERate*** | ***Acc*** | ***SERate*** | ***Acc*** | ***SERate*** |
| AGF | 73 | 0 | 76 | 0 | 67.5 | 0 | 66 | 0 | 71 | 0 |
| BT6 | 68.5 | 0 | 67 | 0 | 71.5 | 0 | 73.5 | 0 | 76.5 | 0 |
| GIL | 57 | 3.5 | 57.5 | 0.5 | 60 | 0.5 | 60 | 0.5 | 60.5 | 1 |
| GMD | 83 | 0 | 79 | 0 | 77.5 | 0 | 81 | 0 | 82 | 0 |
| HAP | 85 | 0 | 86 | 0 | 87 | 0 | 88 | 0 | 85.5 | 0 |
| KHA | 59.5 | 0.5 | 57 | 0 | 63.5 | 0 | 66.5 | 0 | 70 | 0 |
| LAF | 75 | 0 | 93 | 0 | 94.5 | 0 | 95 | 0 | 95.5 | 0 |
| REE | 73.5 | 1 | 84.5 | 1 | 85 | 0 | 85.5 | 0 | 85 | 0 |
| SAM | 74.5 | 3 | 77.5 | 0.5 | 65.5 | 7 | 71.5 | 2 | 75.5 | 0.5 |
| TRI | 86.5 | 0 | 86.5 | 0 | 89.5 | 0 | 90 | 0 | 90.5 | 0 |
| VNINDEX | 88.5 | 0 | 91 | 0 | 88 | 0 | 88 | 0 | 88.5 | 0 |
| ***Trung bình*** | ***74.9*** | ***0.73*** | ***77.7*** | ***0.18*** | ***77.2*** | ***0.68*** | ***78.6*** | ***0.23*** | ***80*** | ***0.14*** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **BPNN** | **DT-BPNN** | **SVM** | **SVM-Prob** | **K-SVMeans** |
| Thời gian (giây) | 15 | 16 | 256 | 251 | 213 |

**Bộ 2**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Mã** | **BPNN** | | **DT-BPNN** | | **SVM** | | **SVM-Prob** | | **K-SVMeans** | |
| ***Acc*** | ***SERate*** | ***Acc*** | ***SERate*** | ***Acc*** | ***SERate*** | ***Acc*** | ***SERate*** | ***Acc*** | ***SERate*** |
| AGF |  |  |  |  |  |  |  |  |  |  |
| BT6 |  |  |  |  |  |  |  |  |  |  |
| GIL |  |  |  |  |  |  |  |  |  |  |
| GMD |  |  |  |  |  |  |  |  |  |  |
| HAP |  |  |  |  |  |  |  |  |  |  |
| KHA |  |  |  |  |  |  |  |  |  |  |
| LAF |  |  |  |  |  |  |  |  |  |  |
| REE |  |  |  |  |  |  |  |  |  |  |
| SAM |  |  |  |  |  |  |  |  |  |  |
| TRI |  |  |  |  |  |  |  |  |  |  |
| VNINDEX |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **BPNN** | **DT-BPNN** | **SVM** | **SVM-Prob** | **K-SVMeans** |
| Thời gian (giây) | 36 |  | 580 | 523 | 403 |

* Dự đoán 5 ngày trên 3 bộ dữ liệu

**Bộ 1**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Mã** | **BPNN** | | **DT-BPNN** | | **SVM** | | **SVM-Prob** | | **K-SVMeans** | |
| ***Acc*** | ***SERate*** | ***Acc*** | ***SERate*** | ***Acc*** | ***SERate*** | ***Acc*** | ***SERate*** | ***Acc*** | ***SERate*** |
| AGF | 77 | 0 | 81.5 | 0 | 66.5 | 0 | 64.5 | 2.5 | 65 | 5 |
| BT6 | 56 | 0 | 52 | 0 | 57 | 0 | 59 | 0 | 66 | 0 |
| GIL | 39 | 21.5 | 40.5 | 7 | 60 | 0 | 59.5 | 1 | 60.5 | 0 |
| GMD | 34.5 | 0 | 34 | 0 | 69 | 0 | 67.5 | 0 | 68 | 0 |
| HAP | 76 | 0 | 74 | 0 | 61 | 0.5 | 65 | 0 | 67.5 | 0 |
| KHA | 51.5 | 4.5 | 52.5 | 0 | 64 | 0 | 67 | 0 | 67 | 0 |
| LAF | 33 | 0 | 90.5 | 0 | 86.5 | 0 | 78.5 | 0 | 87.5 | 0 |
| REE | 71 | 4.5 | 69.5 | 1 | 70.5 | 0 | 74.5 | 0 | 76 | 0 |
| SAM | 48.5 | 22.5 | 63 | 14 | 51.5 | 21 | 56 | 18 | 61 | 13 |
| TRI | 73 | 0 | 61.5 | 0 | 67.5 | 0 | 69.5 | 0.5 | 79 | 0 |
| VNINDEX | 78 | 0 | 81 | 0 | 78.5 | 6.5 | 77 | 1.5 | 78.5 | 1 |
| ***Trung bình*** | ***58*** | ***4.82*** | ***63.6*** | ***2*** | ***66.5*** | ***2.54*** | ***67.1*** | ***2.13*** | ***70.5*** | ***1.73*** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **BPNN** | **DT-BPNN** | **SVM** | **SVM-Prob** | **K-SVMeans** |
| Thời gian (giây) | 15 | 17 | 262 | 249 | 221 |

**Bộ 2**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Mã** | **BPNN** | | **DT-BPNN** | | **SVM** | | **SVM-Prob** | | **K-SVMeans** | |
| ***Acc*** | ***SERate*** | ***Acc*** | ***SERate*** | ***Acc*** | ***SERate*** | ***Acc*** | ***SERate*** | ***Acc*** | ***SERate*** |
| AGF |  |  |  |  |  |  |  |  |  |  |
| BT6 |  |  |  |  |  |  |  |  |  |  |
| GIL |  |  |  |  |  |  |  |  |  |  |
| GMD |  |  |  |  |  |  |  |  |  |  |
| HAP |  |  |  |  |  |  |  |  |  |  |
| KHA |  |  |  |  |  |  |  |  |  |  |
| LAF |  |  |  |  |  |  |  |  |  |  |
| REE |  |  |  |  |  |  |  |  |  |  |
| SAM |  |  |  |  |  |  |  |  |  |  |
| TRI |  |  |  |  |  |  |  |  |  |  |
| VNINDEX |  |  |  |  |  |  |  |  |  |  |
| ***Trung bình*** |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **BPNN** | **DT-BPNN** | **SVM** | **SVM-Prob** | **K-SVMeans** |
| Thời gian (giây) | 37 |  | 596 | 531 | 412 |

* Nhận xét:
* Về hiệu quả giữa các phương pháp:
  + Trong kết quả dự đoán 1 ngày và 5 ngày, nhìn chung KSVMeans cho hiệu quả cao nhất so với BPNN, DT-BPNN, SVM và SVM-Prob xét về độ chính xác lẫn độ lỗi.
  + BPNN vẫn cho thấy hạn chế của nó về khả năng tổng quát hóa dữ liệu khi hiệu suất dự đoán thấp nhất cùng với độ tin cậy kém khi có độ lỗi cao.
  + SVM-Prob tỏ ra nhỉnh hơn SVM truyền thống. Chính vì điều này nên khi cải tiến SVM-Prob bằng cách áp dụng thêm K-Means, K-SVMeans cho kết quả tốt hơn.
  + Sự cải tiến DT-BPNN chưa thực sự ổn định.
* Về độ dài dữ liệu quá khứ:
  + Dữ liệu quá khư dùng huấn luyện ở bộ 1 cho kết quả tốt. Tuy nhiên, khi tăng dữ liệu huấn luyện từ 3 năm (ở bộ 1) lên 4 năm (ở bộ 2), kết quả cải thiện đáng kể. Độ chính xác tăng từ 2-5 điểm phần trăm, kèm theo đó là độ lỗi cũng giảm.
* Về thời gian huấn luyện:
  + BPNN cùng DT-BPNN huấn luyện nhanh nhất do không phải chay thuật toán tìm bộ tham số.
  + Các thuật toán dựa trên SVM cho kết quả lâu hơn, nhưng khoảng thời gian đó là con số chấp nhân được.
    - Bộ 1: 3-5 phút
    - Bộ 2: 8-10 phút
    - Bộ 3: 20-30 phút
  + Trong các phương pháp SVM, SVM-Prob huấn luyện nhanh hơn một ít so với SVM chuẩn và K-SVMeans nhanh hơn SVM-Prob. Sự chênh lệch này khi dữ liệu lớn là khá đáng kể, dễ nhận thấy ở bộ 2 va 3.

# KẾT LUẬN

* Các cải tiến cho BPNN và SVM đều mang lại hiệu quả tốt
* Việc cải tiến dùng Kmeans cho SVM cho kết quả tốt nhất, trong khi DT-BPNN cũng cải thiện hiệu suất so với BPNN tuy chưa ổn định.
* Hiệu quả của DT-BPNN có thể do ảnh hưởng của việc rời rạc hóa các giá trị đầu vào, do đó, cần cải thiện bước xử lý này tốt hơn.
* Về dữ liệu huấn luyện, khoảng thời gian 4 năm quá khứ là phù hợp nhất khi cho kết quả tốt nhất.

# CÔNG VIỆC TƯƠNG LAI