A Hybrid deep network approach for predictive analysis of massive and incomplete data of assurance

Said Achchab
Al Qualsadi research team
ENSIAS, Mohammed V University of Rabat, Morocco
Plan

- Case study:
  - Neural network
  - Bankruptcy prediction
- Proposed method
- Experiments
- Conclusion
Inspired by the neurobiology, neural network is a collection of interconnected neurons that incrementally learn from their environment (data) to capture essential linear and nonlinear trends in a complex data, so that provides reliable predictions for new situations [1].

It’s a method which improves the neural network's performance by updating the weights when a network is simulated in a specific data environment. It can be:

- **Supervised**: the algorithm tries to estimate the dependencies between inputs and targets given.
- **Unsupervised**: the objective is to describe how the data is organized and to extract homogeneous subset without providing a training set.
What is Neural Network?

Some areas of application:

- Biology
- Finance
- Voice recognition
- E-business
- It is the art of predicting bankruptcy and various measures of financial distress of firms
- It is a crucial key indicator for creditors and investors to evaluate the likelihood that a firm may go bankrupt.
Case study: Bankruptcy prediction and risk scoring

**Research problem**

- **Macroeconomic context**
  - Impact small size moroccan firms

- **Firms follow different strategies of declining**
  - Models based on variables measured over one period are not accurate

- **Research problem**
  - **no universally agreed financial ratios list to use** in financial performance prediction models. Missing data is available for small size firms

- **No monitoring tools to scale risk failure for Moroccan firms**
Proposed method: Hybrid Discriminant Neural Networks Model

**Approach**

- Financial ratios selection model
- Predict failure model
- Failure risk monitoring

Discriminant Analysis (DA)  
Backpropagation neural networks (BPNN)  
Self Organizing Maps (SOM)
Proposed method: Hybrid Discriminant Neural Networks Model

Approach

Financial ratios selection model

Predict failure model

Failure risk monitoring

1Y Var

2Y Var

3Y Var
Proposed method: Hybrid Discriminant Neural Networks Model

Financial ratios selection model

Define appropriate variables:

- As there is no theoretical method defining the best input variables of a neural network model, the Discriminant Analysis gives a statistical support in selecting the most relevant subset input vector for the designed neural network model [1].
- Each year before the failure has its own relevant. Therefore, this model selects three subsets of appropriate variables for each year of the period study.
- The results of this step are the input variables for the second layer of the hybrid model.

Predict failure model

- We are on a supervised learning problem for classification.
- Neural Networks with one hidden layer is the best structure to use for classification problems [1] and the back-propagation algorithm is one of the most applied methods to feed-forward networks.

Three BPNN modeling failure predictions respectively for 1, 2 and 3 years before bankruptcy. Each network has its own architecture:

- Input is respectively the preselected discriminate features
- The number of hidden neurons is chosen by experience. Tests will done to choose the number of neurons that gives the minimum of Mean Squared Error
- Output is one binary variable (0 for default and 1 for non-default firms).

Proposed method: Hybrid Discriminant Neural Networks Model

**Learning Algorithm:**

\[ w_{ji}(n) = w_{ji}(n - 1) + \Delta w_{ji}(n) \]
Based on unsupervised learning method, this layer creates a visual representation of firms clustering depending on their risk failure behavior during three years.

Self Organizing Map (SOM) has the ability to project multidimensional data onto a less dimensional data, is an interesting model to define an intermediate risk classes between failed and healthy targets.

The first layer of the network has three inputs representing the firms’ probability of failure during three years.

The output layer is a one dimensional network 1x5 map to define intermediate risk failure levels (1: Healthy, 2: Probably healthy, 3: Moderate failure risk, 4: High risk failure, 5: Very high risk failure).
Proposed method: Hybrid Discriminant Neural Networks Model

- training process:
  - Step 1: initialization
  - Step 2: activation and similarity matching.

Applying the input vector \( X \), and find the winner-neuron \( k \) from \( m \) output neurons using the minimum distance Euclidean criterion:

\[
||x - w_k|| = \min\{||x - w_i|| \, for \, i = 1, ..., m\}
\]

- Step 3: Learning.

Update all weight vectors of all neurons \( i \) in the neighborhood:

\[
w_i(t) := w_i(t-1) + \beta(t)NS(d, t)(x(t) - w_i(t-1))
\]

- Step 4: Iteration.

If \( t > T \) stop, else increment \( t \) and go back to Step 2.
The database used in this research contains a three years period before failure of annual financial statements raw data for a sample of Moroccan firms (127 raw data).

Different samples are collected over the period from 2012 and to 2014.

The firms are balanced 50% healthy and 50% failed by sample and by sector. The database collected contains 933 companies. A binary variable is created with two values (0 if it is failed and 1 if it is healthy).

Data was normalized to bound data values to -1 and +1.

---

**Proposed method:** Hybrid Discriminant Neural Networks Model
Experiments

Financial ratios selection model

- 17, 14 and 18 significant predictor variables are selected respectively for 1, 2 and 3 years before failure models.
- The results, revealed that the average of correct classification rate through 3 years is 70.3% and that decease as the time horizon increases.
- The model output is the subsets selected as appropriate variables to predict firms’ failure in the three time horizons.

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>Actual Class</th>
<th>N-3</th>
<th>N-2</th>
<th>N-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Failed</td>
<td>Healthy</td>
<td>Failed</td>
<td>Healthy</td>
</tr>
<tr>
<td>Failed</td>
<td>442 (94,0%)</td>
<td>243 (52,5%)</td>
<td>444 (94,5%)</td>
<td>250 (54,0%)</td>
</tr>
<tr>
<td>Healthy</td>
<td>28 (6%)</td>
<td>220 (47,5%)</td>
<td>26 (5,5%)</td>
<td>213 (46,0%)</td>
</tr>
</tbody>
</table>

AVG correct classification
- N-3: 70,8%
- N-2: 70,2%
- N-1: 71,8%
Experiments

To fix the architecture of the three BPNN, several test experiences were done:

- Each year, 30 BPNN structures were tested having a number of hidden nodes from 11 to 40 and trained with several learning rates values (0.002, 0.004 and 0.006).
- The convergence criteria used for training are a mean squared error (MSE) less or equal to 0.00001 or a maximum iterations equal to 3000.
- The BPNN topology with minimum MSE is considered as the optimal one.
Experiments

- As results,
  - the topologies chosen are 11, 13 and 12 hidden nodes for respectively BPNN 1Y, BPNN 2Y, BPNN 3Y with 0.004 learning rate.
  - a 4-fold cross validation technique is used to train and test the model to avoid over fitting.
  - This model gives a higher accuracy than DA with 82.7% as average of correct classification rate through 3 years

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>N-3</th>
<th>N-2</th>
<th>N-1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Predicted class</strong></td>
<td><strong>Failed</strong></td>
<td><strong>Healthy</strong></td>
<td><strong>Failed</strong></td>
</tr>
<tr>
<td>Failed</td>
<td>351 (74.4%)</td>
<td>47 (10.1%)</td>
<td>385 (82.0%)</td>
</tr>
<tr>
<td>Healthy</td>
<td>119 (25.3%)</td>
<td>416 (89.9%)</td>
<td>85 (18.0%)</td>
</tr>
<tr>
<td>AVG correct classification</td>
<td>81.4%</td>
<td>82.3%</td>
<td>84.5%</td>
</tr>
</tbody>
</table>
In order to test the hypothesis that a model trained with appropriate variables gives better results than with commonly used variables specially with missing data, same process to build predict failure model is done with a ratios commonly used in literature.

- 5 financial ratios are calculated (Working capital/Total assets, Retained Earnings/Total assets, Earnings before interest and taxes/Total assets, Market value equity/Book value of total debt, Sales/Total assets).
- To define the 3 topologies of BPNN, the same test experiences were done and the results confirm the hypothesis with **81.2%**

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>Actual Class</th>
<th>N-3</th>
<th>N-2</th>
<th>N-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Failed</td>
<td>375 (80.1%)</td>
<td>392 (80.8%)</td>
<td>386 (82.1%)</td>
</tr>
<tr>
<td></td>
<td>Healthy</td>
<td>95 (20.4%)</td>
<td>78 (17.4%)</td>
<td>84 (18.1%)</td>
</tr>
<tr>
<td>Failed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthy</td>
<td></td>
<td>93 (19.9%)</td>
<td>370 (79.6%)</td>
<td>84 (17.9%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AVG correct classification</td>
<td>79.8%</td>
<td>81.7%</td>
<td>82.0%</td>
<td></td>
</tr>
</tbody>
</table>
Experiments

Failure risk monitoring

- we collect the probability of failure to analyze firms’ behavior and simulate risk scoring
- a one dimensional 1x5 map is created
- Group 5, with 360 failed firms, represents 90% of failure prediction accuracy in the same group and 76.5% in the total failed firms
This Chart presents a prototype of firms’ behavior during three years to predict the risk of failure. It confirms the hypothesis that some firms go bankrupt quickly even though they appear healthy (G4); others still survive even if their indicators are alarming (G2 and G3).
Conclusion and perspectives

- This hybrid model takes into account the way firms move in failure space through a period of three years and the constraints of missing data to define their risk failure.
- Based on results, the HDNN gives a good accuracy in comparison with DA.
- It confirms the hypothesis that a model trained with appropriate variables gives better results than with commonly used variables specially with missing data.
- This hybrid model can be a useful tool for investors and stakeholders to define the risk profile of a firms’ portfolio.

Perspectives

- Analyse the significance of discriminante variables selected in the first layer.
- Analyse results by sector.
- Perform the Back propagation learning algorithm and compare it with genetic algorithm.
Thank You