Reviewer’s report

Title: Prediction of Pneumoconiosis Risk and Identification of High Risk Group in Coal Miners Based on Occupational Histories by Neural Network: a retrospective cohort study

Version: 1 Date: 22 June 2009

Reviewer: Mattias Ohlsson

Reviewer’s report:

General

This paper deals with the prediction of Coal workers' pneumoconiosis (CWP) from a single colliery in China, using neural networks as the machine learning tool. It is a retrospective study. Part of the aim for this study is to identify important factors for CWP among the factors used in the classification. The results show that CWP can be accurately predicted and that "dust-exposed duration" and "occupational category" are the most important factors.

The results are impressing, however, the methodology needs clarifications and improvements.

Major Compulsory Revisions:

1. It is not clear why such a complex (15 hidden nodes) neural network model was needed. In section "Development of the neural network" it is stated that between 6 and 15 hidden nodes was tested and that "When there were 15 neurons, the performance of the trained neural network on output sample started to deteriorate."

On what dataset is this performance measured? Clarifications are needed.

2. It is stated that the neural network model had 6 inputs, corresponding to the variables given i Table 1. However the "Job" variable for instance is a categorical variable with 4 categories which needs four inputs following the standard scheme of encoding categorical variables.

Why is there only 6 inputs? How are the categorical variables encoded? If only 6 variables are used then a justification is needed.
3. Regarding the sensitivity analysis: There are many different methods that one can use to determine the sensitivity of the different input features. All methods give different results! The authors should test other methods in order to be more confident in what features are the most important.

Regarding the present approach: According to the paper: "The basic idea of this method is that one input variable to the network is varied within a reasonable interval, while all others are fixed at their respective means,..." What does a reasonable interval means? How much does the result depends on the size of this interval. How are the categorical input variables handled? This last question is also connected to the way you encode the categorical variables. The questions should be clarified in the text.

4. Although the data set is large (14655) there are very few positive cases (236) which makes this a very unbalanced data set. Normally for large data set as this one a single large test set is often sufficient to obtain a reliable performance estimation. However, since the very few positive cases it is reasonable to ask for an improved estimation of the test performance. The authors should at least give a good explanation why they did not use e.g. K-fold cross validation to estimate the test performance.

Minor Essential Revisions

1. There is no justification for the use of 4 significant digits when presenting the sensitivity values. Two digits are sufficient!

   67.01% -> 67%
   64.64% -> 65%

2. Background section.

   Sentence "It can overcome some difficult problems such as local trapping, over-fitting, and overtime in training, et al., ..." I guess the authors mean "... and overtime in training, etc...". Also what does overtime in training mean?

3. The authors should recheck the text for grammatical errors and improvements.

Discretionary Revisions
1. A very common approach when using neural networks is to use an ensemble approach to increase the performance of the network. The authors should perhaps consider to test ensembles to see if an additional performance increase can be obtained.

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Needs some language corrections before being published

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I declare that I have no competing interests.