Accurate Modelling with Minimised Data Collection — An Active Learning Algorithm

Tirthankar RayChaudhuri and Leonard G.C. Hamey

tirthank@mpce.mq.edu.au
len@mpce.mq.edu.au
School of MPCE, Macquarie University, New South Wales 2109, Australia

ABSTRACT

A data gathering method based on active querying is described. In this method data is reduced to a minimum, yet modelling accuracy is uncompromised. Our active querying criterion is determined by whether or not several neural network models agree when they are fitted to random subsamples of a small amount of collected data. Experiments have established the feasibility of our algorithm. It is also shown that our approach results in a more samples being collected in the neighbourhood of the more significant inputs.

1. Introduction: Motivation for Active Learning Research

Minimisation of both generalisation error and data sampling are, ideally, desirable for an efficient learning algorithm. At first glance this appears to be a contradictory pair of objectives. However, just as the generalisation error must be as low as possible for accurate modelling, data sampling which involves both collection and measurement of data is expensive and therefore needs to be reduced to a minimum.

Learning by active querying signifies a learner having the ability to select its own training data by means of queries to the environment. Sollich [12] has used Bayesian analysis, i.e., assumed availability of prior knowledge of a learning task in order to provide justification for active querying. This supersedes the traditional approach of studying generalisation through random examples. It has been shown that random examples contain progressively less information as learning proceeds [7, 13]. Statistical active querying enables the learner to set up criteria in order to select data with maximal information about the system being learned. The quest for more reliable learning techniques has led researchers to examine statistical active querying as a means of obtaining training data that will produce improved generalisation [6]. Such a query-based training process is often called ‘active learning’.

2. A Data Minimising Approach

As already mentioned data gathering is expensive, but computation in today’s world is fast and cheap. Based upon this philosophical premise we suggest a means of achieving both the above goals of active learning simultaneously [9, 8, 10]. We have used the ‘query-by-committee’ approach [3, 11]. If several models are fitted to random subsamples of a small amount of initially collected data, they will probably disagree in the first instance. If we add minimal additional data according to some defined criterion to our sample and repeat the process of having several models examine random subsamples, then after several iterations of this process the models must agree closely at some stage. Our ideas are based upon active learning concepts introduced by Cohn at al [1, 2] and Krogh and Vedelsby [5]. The emphasis of our algorithm however is upon minimising data gathering — without having to compromise modelling accuracy, i.e., increasing generalisation error. Instead of adding one labeled point at a time and having all the data examined by all the neural networks in the ensemble (or ‘committee’) as in Krogh and Vedelsby’s experiments, we have fitted several neural network models with random subsamples of an initially-gathered small data set. The significance of random subsampling is substantiated in Section 4. The querying criterion for resampling is the level of agreement among the models. We continue to resample additional labeled points as long as the models do not agree sufficiently (see Figure 1).

3. Experiments with Clean and Noisy Data

In [9] the feasibility of the proposed scheme has been experimentally established by using feedforward neural network models. A brief description of
the experimental procedure is as follows. A nonlinear system at three different noise levels was examined. The first version (Figure 2(a)) contained clean data and was generated with a feedforward neural network having three hidden nodes and a single node at the input and output respectively. The plots show the system transfer function with the input values along the x-axis and the corresponding outputs along the y-co-ordinate. The second (Figure 2(b)) consisted of the first system with noise added to the output. The third (Figure 2(c)) was a highly noisy form of the first system.

Active learning was performed as hereafter described. It is reasonable to assume that a functional model similar to the one used to generate artificial data can be successfully used to learn from that data. With this premise, each of ten different neural networks, having the same architecture as the one used to generate the clean data, was trained upon half the data (selected randomly) from an initial data sample of ten labeled points, also sampled randomly. The networks were initialised with different sets of random weights. The variance of outputs $y$ of the ten models indicates the level of model disagreement. Expressed formally, this disagreement criterion is thus of the form

$$\frac{1}{n-1} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2$$

the value of $n$ in this case being clearly 10.

This disagreement was computed for 100 randomly chosen values of the input variable $x$. The points of relatively higher variance were those at which more information was needed for correct modelling. It follows therefore that resampling of a labeled point corresponding to maximum variance is desirable. The resampled point was added to the earlier data sample of ten points and the previous process of training the ten networks with random subsamples of half (rounded to the floor integral value) the data was repeated.

The above process was iterated until the maximum model variance dropped below a set threshold — which was 0.001 in this case. Figure 3 shows a set of data points for correct modelling found by active experimentation.

Passive learning was also performed using this algorithm. The difference between active and passive
learning is that the resampling is unbiased in the case of passive learning. Instead of labeled data corresponding to the maximum variance, a random labeled point was resampled from the original system.

For each of the sets of clean, noisy and highly noisy data, 20 active learning and 20 passive learning experiments, were performed — a total of 120 experiments. The histograms of collected data in Figures 4, 5 and 6 show the outcome of the experimental investigations using our algorithm. The histogram bins are simply a range of the number of collected samples for learning in each case. The frequency shows the distribution over these ranges.

4. Higher Sampling Density around Significant Inputs

Krogh and Vedelsby’s investigations [5] did not lead to the definition of a stopping criterion, nor was an analysis made of the amount of data collected. The studies reported in [5] concentrated upon the difference in generalisation error in the cases of active and passive learning. In this work we have added a stopping criterion to the resampling — the threshold value of the maximum model variance. We have have also ensured the existence of model disagreement by our approach of randomly subsampling half the collected data to train each network of the committee, thereby resulting in an adequate number of data points being collected. In [5] all the collected data was given to all the nets at each iteration. In such a case all the nets of the committee will agree as soon as sufficient data points have been collected to define the weight and bias parameters of each network. Although overlap exists between subsamples our random subsampling approach ensures network disagreement more definitely because different networks of the committee essentially examine different parts of the sampled data-space.

Definitive experiments have been performed to illustrate the above claims. We first used Krogh and Vedelsby’s approach to perform active learning
experiments [5] using a square wave step (see Figure 8) for the system to be identified and five neural nets as the identifying ensemble. The nets had 1 hidden node each, i.e., a simple 1-1-1 architecture. The training data began with one labeled point selected randomly from the system. The point corresponding to highest “ensemble ambiguity” [5] or model variance at 1500 random points was resampled in a manner similar to our earlier experiments. The difference between this approach and ours lies in the fact that we randomly subsample half the currently sampled data each time a net is to be trained, whereas in the present method all the networks are trained upon the entire set of collected data samples in each iteration.

Our random subsampling method was also used to identify the same square wave step. For fair comparison purposes, all experimental parameters such as number and type of networks, training algorithm (Levenberg-Marquardt), number of random points for computing variance (1500) — were identical with those of the previous experiments that used Krogh and Vedelsby’s approach. The Matlab Neural Network Toolbox was used for our simulations. Sampled data began with two labeled points selected randomly. A labeled point corresponding to maximum model variance was added at each iteration to the sampled set, half of which was subsampled randomly to train the nets every time.

When Krogh and Vedelsby’s approach was used to identify the square wave step the process converged after the collection of 4 data points. These 4 points defined the two weight and two bias parameters of each network in the committee. Using the random subsampling approach many more points (30) were collected.

In Figure 9 the sequence of data points collected by using our random subsampling approach is shown. Figure 9(a) shows the actual sequence of sampled inputs (the Y-axis being the time sequence) while Figure 9(b) shows the distribution of collected data around each input in the form of a histogram. Clearly the random subsampling approach results in a higher amount of sampled data in the vicinity of each significant input point, i.e., the sharp transition points of the square wave step (Figure 8).

5. Concluding Discussion

It is apparent that the algorithm works in the case of both clean and noisy data — but more points need to be collected for identifying noisy data. Although random resampling (or passive learning) increases computation and results in more samples being collected, it still causes the algorithm to converge.

The advantage of active learning over passive in terms of the number of labeled data points sampled (Figure 7), tends to be less apparent at higher noise levels. In fact in our experiments with the highly noisy system the performance of passive learning was superior. This is interesting and the subject of ongoing investigation.

Finally, using random subsampling of the current sampled data set for training each network is found to result in a higher amount of collected data around each significant input.

6. Acknowledgement

The authors would like to thank Arnott’s Biscuits Ltd for their continuing support in this research.

References


