Noise and Uncertainty Management in Intelligent Data Modeling

Xiaohui Liu and **Gongxian Cheng**
Birkbeck College
Department of Computer Science
University of London, Malet Street
London WC1E 7HX, United Kingdom
hui@dcs.bbk.ac.uk; ubacr46@dcs.bbk.ac.uk

John Xingwang Wu
Institute of Ophthalmology
Department of Preventive Ophthalmology
University of London, Bath Street
London EC1V 9EL, United Kingdom
smgxjow@ucl.ac.uk

Abstract

The management of uncertain and noisy data plays an important role in many problem solving tasks. One traditional approach is to quantify the magnitude of noise or uncertainty in the data and to take this information into account when using this type of data for different purposes. In this paper we propose an alternative way of handling uncertain and noisy data. In particular, noise in the data is positively identified and deleted so that quality data can be obtained. Using the assumption that interesting properties in data are more stable than the noise, we propose a general strategy which involves machine learning from data and domain knowledge. This strategy has been shown to provide a satisfactory way of locating and rejecting noise in large quantities of visual field test data, crucial for the diagnosis of a variety of blinding diseases.

Introduction

Much research has been done to see how real world data can be intelligently modeled using AI methods to produce useful knowledge (Frawley, Piatetsky-Shapiro, & Matheus 1991; Weiss & Kulikowski 1991). Notable examples include the TDIDT (Top Down Induction of Decision Trees) family of learning systems where classification rules are learned from a set of training examples (Quinlan 1986; Bratko & Kononenko 1987). The data are also modeled and directly used to solve problems in application domains. For example, visual field test data are directly used to train neural networks which would associate these data with different kinds of blinding diseases (Nagata, Kani, & Sugiyama 1991).

The real world data, collected or generated in a variety of different environments, however, often contain noise, and are incomplete and uncertain. One of the most challenging research issues in intelligent data analysis is, therefore, how to handle noise and uncertainty in the data so that these data can be used correctly and most effectively in achieving the above described objectives.

One of the traditional approaches to the management of noisy and uncertain data is to use mathematical and statistical techniques to quantify their magnitude in the data and to present general information about the data quality. The decision-making or problem solving process using this type of uncertain information, however, is ultimately a subjective one, depending on one's experience and knowledge. The outcome from this process, therefore, would be often uncertain as well. Also, knowledge discovered from this type of data might be of questionable validity.

In this paper we propose an alternative way of handling noisy data, which has great potential in improving the quality of problem solving and knowledge acquired from data. Instead of measuring and providing information on the amount of noise in the data, we try to explicitly identify and then discard the noise before these data are used for any purpose.

In section 2, the type of noise considered in this paper is defined and a general strategy for its identification is proposed, which involves machine learning from data and domain knowledge. In section 3, this strategy is applied to large quantities of visual field test data which are crucial for the diagnosis of a variety of blinding diseases. In section 4, this strategy is evaluated and we show that it provides a satisfactory way of locating and rejecting noise in the test data. Finally, the work is summarized in section 5.

Noise and its Identification

Measurement Noise

In learning classificatory knowledge from data, there is a universe of objects that are described in terms of a collection of attributes (Quinlan 1986). The objective is to extract from a set of training examples, rules for classifying objects into a number of prespecified categories using those attributes. In these learning systems, data are defined as *noisy* when either the values of attributes or classes contain errors.

In this paper we shall put an emphasis on the errors of attribute values as we are considering the use of data for general purpose applications, not limited to learning classification rules. One of the main reasons for these errors is that the attributes used to describe an object are often based on *measurements*. To illustrate

the idea, consider the task of diagnosing blinding diseases such as glaucoma. A dominating attribute would be to test the visual field of a patient. It is highly unlikely that one could obtain absolutely correct visual field data because these data, collected from patients' responses to visual stimuli on a computer screen, necessarily contain errors caused by various behavioral factors such as the learning effect, inattention, failure of fixation, fatigue etc. These errors are typically in the form of false positive or negative responses from patients (Lieberman & Drake 1992). Quinlan has also given an example of false positive or negative readings for the presence of some substance in the blood (Quinlan 1986).

The *noise* in data considered in this paper refers to incorrect data items caused by measurements. Consequently we shall use the term *measurement noise* throughout the paper.

Identifying the Measurement Noise

One fundamental assumption made in (Becker & Hinton 1992), where a new self-organizing neural network is proposed, is that interesting properties in data are more stable than the noise (Mitchison & Durbin 1992). For example, the property that a normal person who does not have any visual function loss should be able to see the stimuli on the test screen most of the time is more stable than the occasional fluctuation in data caused by errors (e.g. false positive or false negative responses) for whatever reasons. We have adopted this assumption as our basic principle for identifying measurement noise, to which we shall refer as the *noise identification principle.*

Suppose that a repeated test is designed where the same measurement is made a fixed number of times and consider the visual test as an example. A normal person might be distracted in the middle of a test, say, for example the fifth of the repeated measurements. This results in poor sensitivity values for, perhaps, most of the locations within the visual field, leading to fluctuation in the data. This type of fluctuation, however, should not affect the overall results of the visual field as she or he should be able to see the stimuli on the screen during most of the other trials in the test. The main task here is to identify the common feature exhibited by most of the trials, i.e., the person can see the stimuli most of the time. The part of the data inconsistent with this feature, i.e. the fifth trial, will then be exposed and consequently suspected as noise.

The question is, then, how to find a computational method capable of detecting interesting features among data. Unsupervised learning algorithms seem to be natural candidates, as they are known to be capable of extracting meaningful features, which reflect the inherent relationships between different parts of the data (Fisher, Pazzani, & Langley 1991). For example, we can use an unsupervised learning algorithm such as self-organizing maps(Kohonen 1989) to let the data self-organize in such way that more stable parts of data are clustered to reflect certain interesting features, while parts of data which are inconsistent with those features will be separated from the stable cluster.

It should be emphasized that the less stable part of data should not necessarily be the measurement noise in that they can be actually the true measurements reflecting real values of an attribute. In the example of diagnosing glaucoma using visual field data, the fluctuation in the data can be caused by behavioral factors such as fatigue and inattention, but can also be caused by pathological conditions of the patient. Consider that a glaucoma patient undergoes a visual field test. It is quite possible that there will be still fluctuations in the responses at certain test locations, even if s/he has fully concentrated during the test. The nature of the disease has dictated her/his responses. The elimination of these responses would lead to the loss of much useful diagnostic information, and worse still, could lead to incorrect conclusion about the patient's pathological status.

Therefore, it would be desirable to check whether the less stable part of data is indeed the measurement noise. This is difficult to achieve using the data alone, as there are often many possible explanations for fluctuation in the same data set, as discussed above. The use of a substantial amount of domain specific knowledge, however, has potentials in resolving this difficulty. For example, the knowledge of how diseases such as glaucoma manifest themselves on the test data is crucial for identifying the measurement noise, as we can then have a better chance of finding out the component within the less stable part of the data, which is caused by pathological reasons.

The above discussions lead to a general strategy for identifying the measurement noise in data, which consists of two steps. Firstly, an unsupervised learning algorithm is used to cluster the more stable part of the data. This algorithm should be able to detect some interesting features among those data. The less stable part of the data, which are inconsistent with those features, then becomes the suspect of measurement noise.

Secondly, knowledge in application domains, together with knowledge about the relationships among data, is used to check whether the less stable part of data is indeed the measurement noise. This type of domain specific knowledge may be acquired from experts, however, it is often incomplete. For example, only a partial understanding has been obtained about how diseases like glaucoma manifest themselves on any visual field test data (Wu 1993). Therefore, it is often desirable to apply machine learning methods to the initially incomplete knowledge in order to generalize over unknown situations. One such example is shown in the next section.

Identifying Noise in Glaucomatous Test Data

The Computer Controlled Video Perimetry (CCVP). The CCVP (Fitzke *et al.* 1989; Wu 1993) is a newly developed visual function test method and has been shown to be an effective way of overcoming difficulties in the early detection of visual impairments caused by glaucoma. It examines the sensitivity of a number of locations in the visual field using vertical bars on the computer screen [see Figure 1 for an example]. All these locations are tested by several different stimuli and the test is repeated a fixed number of times. One popular version of the CCVP test examines 6 locations using the same stimulus and the test is repeated 10 times.

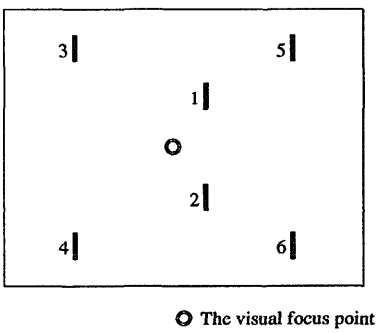

Figure 1: A CCVP screen layout

If the stimulus is seen at any stage of the test, the patient presses a button as a response. At the end of this CCVP test, ten data vectors are produced, each of which records the patient's response during a single trial. Each vector consists of 6 data elements referring to the results of testing 6 locations using the same stimulus. As far as each location is concerned, there will be a sensitivity value calculated by counting the percentage of positive responses. The clinician relies heavily on these location sensitivity values to perform diagnosis.

Applying the Strategy to the CCVP Data

Identifying the More Stable Part of the CCVP data. The method for identifying the more stable part of the CCVP data is to model the patient's test behavior using the self-organizing maps (SOM). Data clusters can then be visualized or calculated. This method consists of three steps.

Firstly, Kohonen's learning technique (Kohonen 1989) is used to train a network capable of generating maps which reflect the patient's test behavior. Each response pattern for each test trial is used as an input vector to the self-organizing map and each winner node is produced on the output map. In all, 2630 trial data vectors corresponding to 263 tests are used to train the network and the whole data set is reiteratively submitted 100 times in random orders.

Secondly, an effort is made to find a network which shows better *neighborhood preservations*, i.e. similar input patterns are mapped onto identical or closely neighboring neurons on the output map. This step is important as we want to map similar response patterns from patients onto similar neurons. We have used the *topographical product* (TP) (Bauer & Pawelzik 1992) as a measurement for this purpose where TP indicates the magnitude of neighborhood violation. Therefore, the smaller the value of TP is, the better the neighborhood preservation would become.

Having obtained a well-performed network, the final step is to generate the behavior maps for individual patients and analyze these maps to identify the more stable part of data. As far as each patient is concerned, there would be ten winner nodes and nine transitions on the output map. These transitions constitute a transition trajectory, which graphically illustrates how patient's behavior changed from one trial to the other [Figure 2].

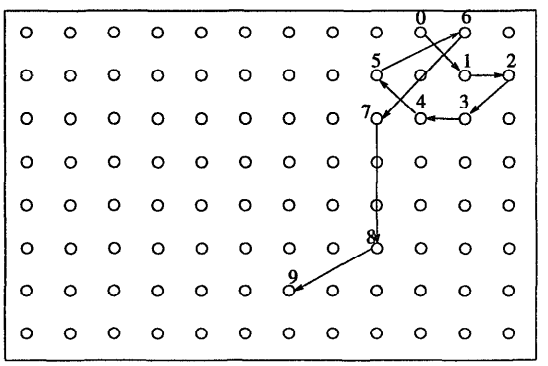

Figure 2: A transition trajectory in the output map

As one of the key SOM features is that similar input vectors would lead to similar winner nodes, here we have the general rule for identifying the more stable part of the data: if most of the winner nodes are centered around one particular region, then the input data vectors associated with these nodes constitute the more stable part of the data. These vectors share one common feature: they are similar to each other, judged to a large extent by a distance measurement such as the Euclidean distance.

The above rule can be implemented by algorithms using the geometry positions of the nodes and their relative distances. The approach taken here is to search for a maximum set of neurons on the output map, which occupies the smallest topographical area. In particular, an evaluation function is defined in equation 1 for this purpose and the objective is to find a subset of winner nodes, S, which minimizes the value of $F(S)$.

$$F(S) = A(S(k))/k^2 \quad (k = N, N-1, ..., \lfloor N/2+1 \rfloor) \quad (1)$$

Where N is the total number of winner nodes (ten

in our application), A denotes the topographical area in the map occupied by a subset of winner nodes, and $S(k)$ represents a subset of winner nodes with k members.

Checking the Less Stable Part of Data. Let us now examine the less stable part of data, for example, the data vectors associated with winner nodes 8 and 9 in Figure 2, and see whether or not some of these vectors are the measurement noise. For our chosen application, we are particularly interested in finding out whether data items within this less stable part of data are caused by pathological conditions of the patient during the visual field test.

To achieve this, a deep understanding of how diseases manifest themselves on the data is essential. Here we have used both knowledge about inherent relationships among data and domain knowledge from experts to obtain this understanding.

The knowledge about data is reflected on the maps produced by the SOM. For example, each neuron on the output map is likely to have a number of input vectors associated with it, and these input vectors in turn determine the physical meanings of the neuron such as average sensitivity, the number of input patterns the neuron represents, and typical patterns the neuron represents etc. Using these physical meanings, domain experts can try to group those input patterns which have the same or similar pathological meanings. In our case, an input pattern consists of a vector of 6 elements, each of which represents whether the patient sees the stimulus in a certain location on the computer screen [Figure 1].

There are four major groups created by experts. Group A is composed of those input patterns reflecting that the patient under test is showing the early sign of upper hemifield damage, while group B consists of those patterns demonstrating that the upper hemifield of the patient is probably already damaged. Group C and D are made of those patterns similar to group A and B, except they are used to represent two different stages of the lower hemifield damage. Any two patterns which fall into the same group, no matter how distant they may appear on the test behavior map, will be considered as having the same pathological meanings.

Take group A as an example. It contains the following three patterns:

$$\{\ (1,1,0,1,1,1)^t,\ (1,1,0,1,0,1)^t,\ (1,1,1,1,0,1)^t\ \}$$

These have been identified as possible patterns for a glaucoma patient showing early sign of upper hemifield damage. Two factors have been taken into consideration by experts when selecting these patterns. Firstly, the domain knowledge about early upper hemifield damage is used, for example, locations 3 and 5 which are within the upper hemifield were not seen in some of those patterns and the reason why location 1 is not included is that it often indicates the upper hemifield is probably already damaged (Wu 1993). Secondly, the physical meanings of the trained map are used, especially how typical input patterns are associated with output neurons. For example, the above three patterns are in a topographically connected area on the map.

These pathological groups are then used to check whether the less stable part of the data are the measurement noise. A simple way to do this is as follows. When those nodes, whose corresponding input data vectors are the less stable part of the data, are identified, check whether each of these data vectors belongs to the same pathological groups as those patterns which were recognized as the more stable part of data. If yes, then treat it as a true measurement; otherwise, it is measurement noise.

One of the major difficulties in applying this method is that the patterns which are made up those pathological groups are not complete in that they (27 in total) are only a subset of all the possible patterns ($2^6 = 64$). Therefore, when there is a new pattern occurring, the above method cannot be applied. One of the main reasons why experts cannot classify all the patterns into those four groups is that the CCVP is a newly introduced test and the reflection of glaucoma patients and suspects on the CCVP data is not fully understood.

To overcome this difficulty, machine learning methods can be applied to generalize from those 27 classification examples provided by the experts. In particular, we have used the back-propagation algorithm (Rumelhart, Hinton, & Williams 1986) for this purpose. The input neurons represent the locations within the visual field, output neurons are those pathological groups, and three hidden nodes are used in the fully configured network.

The trained network is able to reach 100% accuracy for the training examples and to further classify another 26 patterns. One of the interesting observations is that patterns within each of the resultant groups tend to be clustered in a topographically connected area, a property demonstrated by the initial groups. The remaining patterns are regarded as the unknown class since they have no significant output signal in output neurons. They have been found to be much more likely to appear in the less stable part of the CCVP data than in the more stable one.

It should be noted that the application described above is rather a simple one in which there are only 64 possible input patterns. This particular version of the CCVP test is chosen for its simplicity in order to make it easier to describe the general ideas in implementing the noise identification principle. In fact, there is a more popular version of CCVP which also tests the six locations within the visual field by ten repeated trials, however, using *four* different stimuli. Therefore the data vectors produced within this test contain 24 items, instead of 6, and consequently, there are 2^{24} possible input patterns. We have also experimented

with large quantities of data from this test using the proposed noise identification strategy. The results are similar to those of the simpler test described in the next section.

Evaluation

The Strategy

The noise identification strategy is based on the assumption that interesting properties in data are more stable than the noise. It can be applied to those areas where repeated measurements can be easily made about attributes concerned. Below are several observations regarding this strategy.

Firstly, explicit identification and deletion of measurement noise in data may be a necessary step before the data can be properly explored, as shown in our application. In particular, we have found that noise deletion can offer great assistance to the clinician in diagnosing those otherwise ambiguous cases (see section 4.2). In a separate experiment with learning hidden features from the CCVP data, we have found that many useful features, such as behavioral relationship between two test locations, were not initially found from the raw CCVP data, but were uncovered from the data after the measurement noise was deleted using the strategy proposed in this paper.

Secondly, the use of domain knowledge supplied by experts is of special concern as this type of knowledge involves a substantial amount of subjective elements, and is often incomplete as shown in our application. It should be pointed out that this strategy cannot be applied to those applications where there is little relevant high quality knowledge but a lot of *false noise*, i.e., those data items from the less stable part of the data which actually reflect the true measurements. Where there is little concern about the false noise situation, however, an unsupervised learning algorithm can be used directly to identify the measurement noise, in this case, the entire less stable part of the data.

Finally, no claim is made that this strategy can be used to identify all the measurement noise in data, or all the noise identified is the real one. This depends on the ability of the chosen algorithms to accurately cluster those data items with common features and the quality of domain knowledge used to exclude the false noise.

The Results

Here we present the results in applying the proposed strategy to a set of clinical test data (2630 data vectors) collected from a group of glaucoma patients and suspects. To find out how successful this strategy is in achieving its objective, we use the idea of *reproducibility* of the test results.

As glaucoma is a long term progressing disease, the visual function should remain more or less the same during a short period of time. Therefore results from such two repeated tests within this time period should be very close. However, this is not always true under real clinical situations as measurement noise is involved in each test, perhaps for different reasons. Thus it is not surprising to note that there are a large number of repeated tests, which were conducted within an average time span of one month, whose results showed disagreements to various degrees.

As one of the main reasons for the disagreement is the measurement noise, it is natural to assume that the sensitivity results of the two tests should agree (to various degrees) after the noise is discarded. This then constitutes a method for evaluating our proposed strategy for identifying and eliminating noise from data.

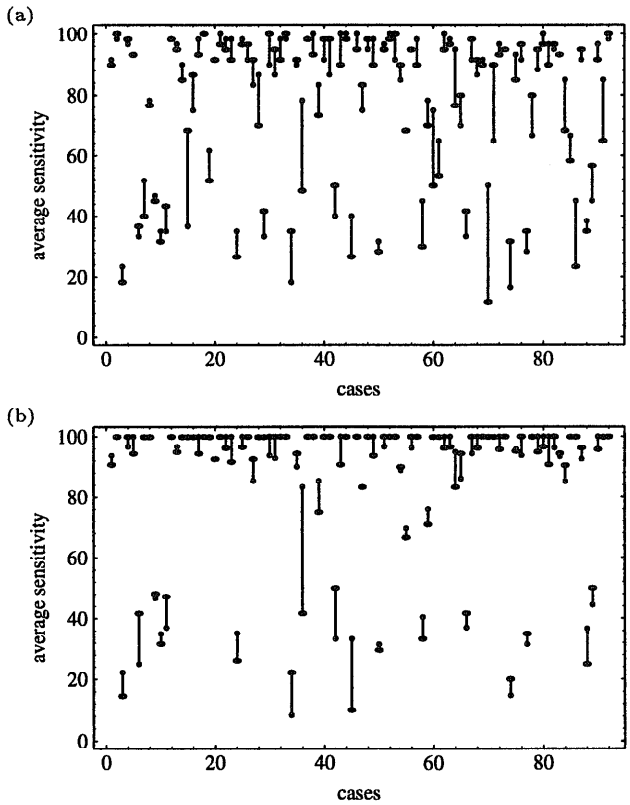

Figure 3: (a) before deletion; (b) after deletion

Ninety-two pairs of test records are available for this purpose. The average sensitivity values of these tests are contrasted in Figure 3(a) where the dot is used to indicate the result of the first test, the oval is used for the result of the second test, and the difference between the two results for each case is illustrated by the line in between them. The same results after the rejection of noise by the proposed strategy are given in Figure 3(b).

The results from the two repeated tests have much better agreements after the noise is rejected. This is indicated by the observation that the lines between the

two tests are in general shortened in Figure 3(b). In fact, if one calculates the mean difference between the two tests, 5.4 is the figure for the original data, while 3.6 is obtained after the noise is eliminated.

Another major finding is that noise deletion may also be of direct diagnostic assistance to the clinician. One of the difficulties for the clinician is that the result from one test suggests that the patient is *normal* (no glaucoma), while the result from the other test shows that the patient is *abnormal* (having glaucoma of some kind). It has been found that the average sensitivity value of 75% appears to be the golden line n CCVP that divides the normal and abnormal groups (Wu 1993). Since much better agreement is shown between the two repeated tests after the deletion of noise, there would be fewer cases whose test results are split by the golden line. This is indeed the case with our data as shown in Figure 3: there are quite a few conflicting cases in Figure 3(a), while only about two such cases exist in Figure 3(b).

It is worth reiterating that the CCVP is a newly introduced test method. A deeper understanding of its characteristics and its relevance to diagnosis can help further improve the results of identifying the measurement noise in the CCVP data.

Concluding Remarks

In this paper we have introduced an alternative way of dealing with noisy data. Instead of measuring and providing information on the amount of noise in the data, we explicitly identify and then discard the noise so that quality data can be used for different applications.

The principle we adopted for identifying measurement noise is that interesting properties in data are more stable than noise. To implement this principle for our application, self-organizing maps are used to model patient's behavior during the visual field test and to separate the more stable part of data from the less stable one. Expert knowledge, augmented by supervised learning techniques, is also used to check whether data items within the less stable part are measurement noise caused by behavioral factors, or those caused by the patient's pathological conditions.

The proposed strategy has been shown to be a satisfactory way of identifying measurement noise in visual field test data. Moreover, the explicit identification and elimination of the noise in these data have been found not just desirable, but essential, if the data are to be properly modeled and explored. Finally, the strategy may be used as a preprocessor to a variety of systems using data with measurement noise.

Acknowledgements

This work is in part supported by the International Glaucoma Society, British Council for Prevention of Blindness, and International Center for Eye Health. We would like to thank Phil Docking for his comments on an early draft of this paper and anonymous referees' informative review.

References

Bauer, H. U., and Pawelzik, K. R. 1992. Quantifying the neighborhood preservation of self-organizing feature maps. *IEEE Trans. on Neural Networks* 3(4):570–9.

Becker, S., and Hinton, G. E. 1992. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature* 355:161–163.

Bratko, I., and Kononenko, I. 1987. Learning diagnostic rules from incomplete and noisy data. In Phelps, B., ed., *Interactions in Artificial Intelligence and Statistical Methods.* Technical. 142–53.

Fisher, D. H.; Pazzani, M. J.; and Langley, P. 1991. *Concept Formation: Knowledge and Experience in Unsupervised Learning.* Morgan Kaufmann.

Fitzke, F. W.; Poinoosawmy, D.; Nagasuberamanian, S.; and Hitchings, R. A. 1989. Peripheral displacement threshold in glaucoma and ocular hypertension. *Perimetry Update 1988/89* 399–405.

Frawley, W. J.; Piatetsky-Shapiro, G.; and Matheus, C. J. 1991. Knowledge discovery in databases: An overview. In Piatetsky-Shapiro, G., and Frawley, W. J., eds., *Knowledge Discovery in Databases.* AAAI Press / The MIT Press. 1–27.

Kohonen, T. 1989. *Self-Organization and Associative Memory.* Springer-Verlag.

Lieberman, M. F., and Drake, M. V. 1992. *Computerized Perimetry.* Slack Inc.

Mitchison, G., and Durbin, R. 1992. Learning from your neighbour. *Nature* 355:112–113.

Nagata, S.; Kani, K.; and Sugiyama, A. 1991. A computer-assisted visual field diagnosis system using a neural network. *Perimetry Update 1990/91* 291–95.

Quinlan, J. R. 1986. Induction of decision trees. *Machine Learning* 1:81–106.

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1986. Learning representations by back-propagating errors. *Nature* 323:533–36.

Weiss, S. M., and Kulikowski, C. A. 1991. *Computer Systems that Learn.* Morgan Kaufmann.

Wu, J. X. 1993. *Visual Screening for Blinding Diseases in the Community Using Computer Controlled Video Perimetry.* Ph.D. Dissertation, University of London.