Concise Papers

Analyzing Outliers Cautiously

Xiaohui Liu, Gongxian Cheng, and John X. Wu

Abstract—Outliers are difficult to handle because some of them can be measurement errors, while others may represent *phenomena of interest*, something "significant" from the viewpoint of the application domain. Statistical and computational methods have been proposed to detect outliers, but further analysis of outliers requires much relevant domain knowledge. In our previous work, we suggested a knowledge-based method for distinguishing between the measurement errors and phenomena of interest by modeling "real measurements"—how measurements should be distributed in an application domain. In this paper, we make this distinction by modeling measurement errors instead. This is a cautious approach to outlier analysis, which has been successfully applied to a medical problem and may find interesting applications in other domains such as science, engineering, finance, and economics.

Index Terms—Outliers, domain knowledge, AI modeling, self-organizing maps.

1 Introduction

A strange data value that stands out because it is not like the rest of the data in some sense is called an outlier. Often, a mistake in the data will result in an outlier being present. However, not all outliers are mistakes and these "unrepresentative" data points may represent phenomena of interest, something "significant" from the viewpoint of the application domain.

There are two principal approaches to outlier management [2]. One is outlier *accommodation*, which is characterized by the development of a variety of statistical estimation or testing procedures which are *robust* against, or relatively unaffected by, outliers [11]. In these procedures, the analysis of the main body of data is the key objective and outliers themselves are not of prime concern.

The other approach, characterized by identifying outliers and deciding whether they should be retained or rejected, is the subject of study in this paper. Many statistical techniques have been proposed to detect outliers and comprehensive texts on this topic are those by Hawkins [10] and Barnet and Lewis [2]. These approaches range from informal methods such as the ordering of multivariate data [1], the use of graphical and pictorial methods [12], and the application of simple test statistics [7] to some more formal approach in which a model for the data is provided and tests of hypotheses that certain observations are outliers are set up against the alternative that they are part of the main body of data [10], [4]. The identification of outliers has also received much attention from the computing community [3], [13], [17]. However, there appears to be much less work on how to decide whether outliers should be retained or rejected. In order to successfully distinguish between noisy outlying data and noise free outliers, different kinds of information are normally needed. These should not only include various data characteristics and the context in which the outliers occur, but also relevant domain knowledge. The procedure for analyzing outliers has been experimentally shown to be subjective, depending on the above mentioned factors [6].

We have conducted some preliminary research on how to analyze outliers using domain knowledge. In particular, a strategy for distinguishing between phenomena of interest and measurement noise was proposed and applied to the analysis of a set of visual field test data collected from a group of glaucoma patients in an eye hospital [16]. That strategy attempted to model "real" measurements, namely, how measurements should be distributed in a domain of interest (e.g., how glaucoma manifests itself on visual field data) and rejected values that do not fall within the real measurements. In this paper, however, we attempt to model noise and error processes instead and accept data outside of the norms if it is not accounted for by a noise model. We describe a program that uses this method which does significantly better at a diagnostic task than an equivalent approach that either utilizes all data or attempts to reject all nonnormal values.

2 Outlier Analysis by Modeling Noisy Data

Assuming that we have a model of how noisy data points are distributed in an application, we could then use that model to help distinguish between the noisy outlying data and the *noise-free* outliers. In this section, the strategy for constructing the noisy model is detailed, followed by the discussion of its application to a medical problem in subsequent sections.

Definition. *Let Ω be a p-dimensional sample space. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of vectors drawn from Ω and O be a set of outliers in X, where $O \subset X$. Let $C = \{noisy, rest\}$ represent two general classes. Let $F = \{f_1, f_2, \ldots, f_m\}$ be a set of features extracted from X.*

Definition: Noise Model Construction. *A noise model is constructed which could account for much of recognized measurement noise in a domain of interest. Here, we assume such a model is not readily available (things would become much easier if it is), and it needs to be constructed or learned. In particular, we assume that a group of data sets Xs can be labeled into two general classes (noisy, rest), based on relevant domain knowledge and close examination of data sets. This group of labeled data sets, together with a set of features F which may be extracted from the data sets, is then used to build the classification model (e.g., a set of classification rules) using an inductive learning technique. Those classification rules corresponding to $noisy$ then become the noise model M.*

Definition: Noise Elimination. *Each new data set, say X', is tested against the noise model M generated in the above step. If applicable, then the outliers O' within the data set X' can be rejected (due to known measurement noise).*

We make the following observations regarding the proposed method:

1. The construction of the noise model requires a set of representative labeled instances on which the "noise model" may be built. There are two mutually exclusive classes that each data set can be assigned to in a domain of interest. Class $noisy$ indicates that the corresponding outliers O in the data set X are noisy data points and can therefore can be deleted; while class $rest$ either says we are not sure whether the outliers in the data set are due to

- *X. Liu is with the Department of Information Systems and Computing, Brunel University, Uxbridge, Middlesex, UB8 3PH, UK. E-mail: Xiaohui.Liu@brunel.ac.uk.*
- *G. Cheng is with the School of Computer Science and Information Systems, Birkbeck College, University of London, Malet St., London WC1E 7HX, UK. E-mail: ubacr46@dcs.bbk.ac.uk.*
- *J.X. Wu is with Moorfields Eye Hospital, Institute of Ophthalmology, Bath St., London EC1V 9EL, UK. E-mail: j.wu@ucl.ac.uk.*

Manuscript received 8 Jan. 1998; revised 16 June 2000; accepted 4 Jan. 2001; posted to Digital Library 10 Sept. 2001.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number 106134.

1041-4347/02/$17.00 © 2002 IEEE

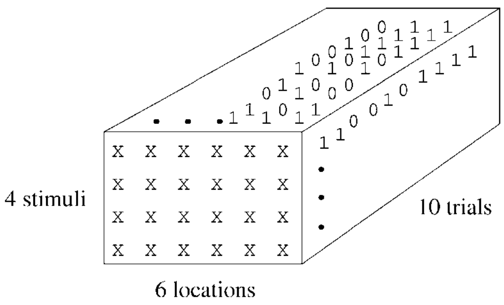

Fig. 1. Data structure of each CCVP test.

measurement noise or some phenomena of interest, or when there is no outlier in the data set.

2. Note that many classification models may be constructed from a set of labeled instances. The classification model as mentioned in the *Noise Model Construction* step refers to the "best" model in terms of its predictive accuracy, its simplicity or interpretability, misclassification costs, or other appropriate criteria for the problem under investigation [5], [8], [19].
3. The success of the method very much depends on the correctness and completeness of the noise model constructed. The correctness of the model depends largely on the quality of domain knowledge—a set of labeled instances in the proposed method—although the choice of inductive learning algorithm may also matter. On the other hand, if the model does not sufficiently cover all possible types of measurement noise, the data set after cleaning would still contain much noise.
4. The precondition for using the method regarding the availability of relevant knowledge about the distribution of noisy data points (labeled instances) is reasonable in many applications. For example, in time series forecasting, the understanding of "special irregular events" and their effects on the forecasting results can be used to build the corresponding "noise model" where using the data after removing the effect of the special irregular effects may often increase the forecasting accuracy. In the next two sections, we will present a detailed case study of analyzing medical test data and show how such noise model can be developed to eliminate the measurement noise.

3 Outliers in Visual Field Data

The Computer Controlled Video Perimetry (CCVP) [20] is a 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 six locations on the test screen which correspond to crucial positions in the visual field. All six locations are tested by one or more stimuli and this measurement cycle is repeated 10 times.

The test screen consists of a number of objects of the same type and, at any stage of the test, only one of them is moving. If the stimulus is seen at any stage of the test, the subject presses a button as a response. At the end of each CCVP test, 10 data vectors are produced, each of which records the subject's responses during a single measurement cycle. The corresponding data structure for each test consisting of 10 repeated measurements over six locations using four stimuli can be seen in Fig. 1 where each element is either *1* which shows that the subject does see the stimulus or *0* which shows the subject does not.

One way of identifying outlier(s) is to use Kohonon's self-organizing map (SOM) [14]. The SOM consists of two layers of nodes: the input layer and the output layer. The input layer is a vector of N nodes for presenting the input patterns to the network, and the output layer is often a two-dimensional array of M output nodes (output map). Each input node is fully connected to every output node via a connection weight, so there is a *weight vector* associated with each output node. When an input vector is presented to the SOM, the distance between it and each of the weight vectors is computed. The output node whose weight vector is closest to the input vector is called the *winner node*.

An SOM is capable of mapping similar input patterns onto geometrically close output nodes. If a majority of the winner nodes can be located in a small neighborhood as a cluster on the output map, then those data vectors that correspond to a few remaining nodes that are far away from this neighborhood are exposed as outliers. The following are the key steps involved in obtaining SOM for the identification of outliers in visual field data [15]:

1. *Definitions.* Let V be the input data set such that each $v^i \in V$, a vector of N dimensionality with values over the set $\{0, 1\}$, corresponds to the response pattern from one measurement cycle such that v^i_k is 1 if the subject can see the *k*th stimulus within the cycle, and 0 otherwise.Let the output space be A and the number of output nodes be M.

   Let W be a set of connection weight vectors where each output node $j (1 \leq j \leq M)$ is associated with a connection weight vector of N dimensionality of the form $w_j = (w_{j1}, \ldots, w_{jN})$, where w_{jk} represents the connection weight between the input node k and the output node j.

   Let η be the gain value which affects the rate of adjustment of the connection weight vectors, and let N_c be a round neighborhood of output node c and r be the radius of N_c. For any output node $j, j \in N_c$ if the distance between c and j in the output map is not greater than r.
2. *Initialize the topology and size of the output map.* Set the value for M.
3. *Initialize weights.* Initialize the connection weights to random values over the interval [0.0, 1.0], and normalize both the input vectors and the connection weight vectors. Initialize the gain value η and the neighborhood radius r.
4. *Present new input.* Set i to i + 1 and present input vector v^i.
5. *Select minimum distance.* The distance between the input vector v^i and each output node j is computed by

   $$d(v^i, w_j) = \sum_{k=1}^{N} (v^i_k - w_{jk})^2.$$

   Designate the winner node with minimum distance to be c.
6. *Update weights and neighborhood.* Adjust the connection weight vectors of the winner node c and its neighborhood N_c, i.e., for each node $j \in N_c$ perform the following:

   $$w_j^{(new)} = w_j^{(old)} + \eta[v_i - w_j^{(old)}].$$

   Decrease both the neighborhood radius r and the gain value η.
7. *Repeat by going to 4.* This iteration process continues until a stable network is obtained. In our experiments, several thousand input vectors were used to construct the network and they were iteratively submitted 100 times in random orders to achieve convergence.

Fig. 2 illustrates the application of the SOM to the CCVP test. The basic idea is that the data set corresponding to each

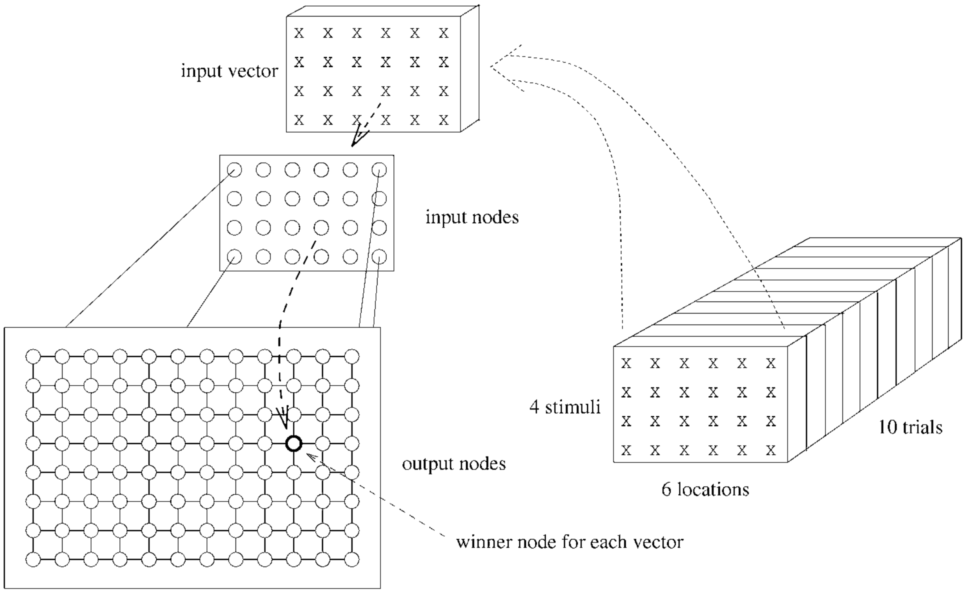

Fig. 2. Apply the SOM to the test.

measurement cycle is used as an input vector and each input vector produces a winner node, so the 10 data vectors for each CCVP test would produce 10 winner nodes.

SOM has an important property that similar data vectors are mapped onto identical or geometrically close *winner* nodes on the output map, making the clustering and visual inspection of similar data vectors possible. Fig. 3 demonstrates the results of a particular test case on the SOM output map. The meanings of the map in the context of visual field data is this: The top right region of the map is most *sensitive*, i.e., the subject can see most of the time and the sensitivity gradually fades toward the bottom left region (see [15] for details). Therefore, it is not difficult to see that this subject could see all the time except during the eighth measurement cycle, in which case the subject may have a normal visual field, but was distracted during that particular measurement cycle. The data vector collected during the cycle is therefore an outlier, while those vectors corresponding to the other nine cycles form the stable part of the data.

Note that there is something arbitrary about any rule that decides which data values are outliers and which are not. Take credit authorization as an example. The rule for identifying suspicious, possible fraud transactions, is not always easy to set. If the rule is too strict, fewer outliers are detected, hence, some important ones (fraudulent transactions) may be missed; if the rule is too lenient, many more outliers (including a lot of legal transactions) are detected, requiring attention to details that are not important. To identify the main cluster of data using SOM, thus exposing outliers is just one of the informal approaches to outlier detection. One of the advantages of this approach is that it can visualize the subject's behavior during the test, e.g., whether the subject has demonstrated the signs of fatigue, inattention, and learning effects, etc. For example, Fig. 3 may have shown a case of inattention by the subject during the eighth test cycle. This visualization property is helpful in labeling the training cases for building the noise models (see the next section).

4 Analyzing Visual Field Outliers

Fig. 4 illustrates how the strategy proposed in Section 2 based on noise models works. Suppose that a set of training data, by using relevant domain knowledge, can be labeled into two classes:“noisy” and “rest.” Class “noisy” indicates that the corresponding outliers in this data set are noisy data points, while class “rest” covers all other situations. Given sufficient amounts of training data, one can use any supervised machine learning techniques to build a “noise model” and this model, after validation, can then be used to help distinguish between the two types of outliers.

Note that the labeling of training instances is not always easy, especially with multidimensional data. To assist in this process, we have used Self-Organizing Maps to visualize and compress data into a two-dimensional map as discussed in Section 3. Data clusters and outliers then become easy to spot, and data are then relatively easily interpreted using the meaning of the map and relevant domain knowledge. So given a data set, outliers may be detected and can then be tested using the noise model. As a result, noisy outliers can then be deleted, while the rest of outliers are kept in the data set for further analysis.

4.1 Noise Model I: Noise Definition

In this application, noise in data are defined as those data points typically associated with *learning effects*, *fatigue*, and *inattention*.

In this connection, we may define outlying data points due to *learning effects* on the self-organizing maps as follows: If the sensitivities of the initial few cycles do not show much regularity, but the sensitivities of the remaining cycles gradually become

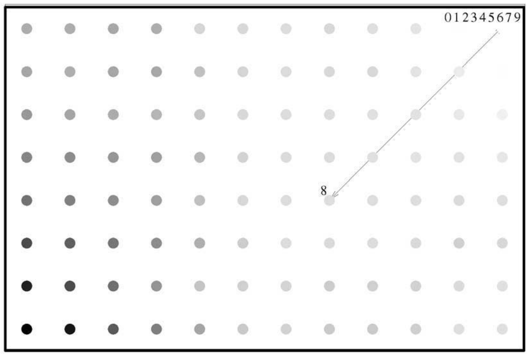

Fig. 3. Visualizing outlier/inattention on the SOM output map.

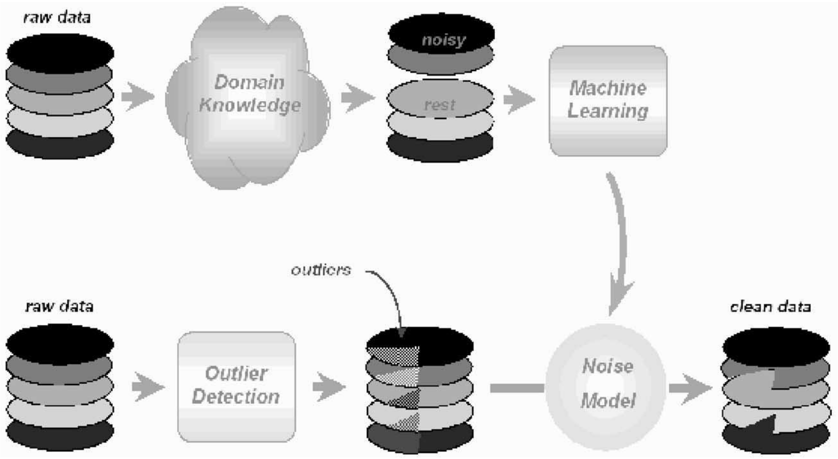

Fig. 4. The process of outlier analysis.

similar, then the data points corresponding to the initial cycles are outlying due to learning effects. In this case, the winner nodes of the initial cycles perhaps are irregular, but are gradually gathered around a small neighborhood on the map. One example of such cases can be seen in Fig. 5.

On the other hand, if the sensitivities of several initial test cycles are high and similar to each other, and the sensitivities of the remaining cycles are decreasing over time, then the data points corresponding to the remaining cycles are outlying due to *fatigue*. In this case, the winner nodes of the initial cycles tend to be in a small neighborhood and the winner nodes of the last few cycles tend to move away from the small neighborhood to areas where the sensitivities are lower. One example of such cases is given in Fig. 6 where the subject could see most of the time during the first six measurements, while the nodes of the next four cycles move away from the first node to some nodes with lower sensitivity values, which indicates that the subject cannot see as clearly as early in the test.

A typical case of *inattention* was given early in Fig. 3 where clearly the subject had a normal visual field, but was distracted during the eighth measurement cycle. This results in poor sensitivity values for this particular measurement cycle, leading to fluctuation in the data. This type of fluctuation, however, should not affect the overall results of the visual field. Therefore, the data collected during this cycle can be dropped. In all the above cases, decisions regarding whether to delete certain outlying data points are relatively easy. For example, the data points corresponding to measurements 6, 7, 8, and 9 may be deleted in Fig. 6, while those corresponding to measurements 0, 1, 2, and 3 in Fig. 5 may be cleared. However, things are not always this clear-cut.

Figs. 7 and 8 demonstrate two test results for the same subject who had been confirmed by an ophthalmologist as a glaucoma patient. Fig. 7 does seem to show there is a cluster in the top left corner of the map. However, since none of the measurements has shown any high sensitivities (in the top right corner area) and there are six measurements scattered around on the map, there is good reason to believe that these measurements might tell us something about the pathological status of the subject. The elimination of any of these measurements might lead to the loss of useful diagnostic information, and worse still, could lead to an incorrect conclusion about the patient's pathological status. Meanwhile, Fig. 8 does not seem to show any interesting clusters and none of the measurements are very sensitive. In this case, there is no easy way of finding out which measurements are noisy and which are not, therefore all the measurements are kept for further analysis.

4.2 Noise Model II: Construction

The construction of the noise model in this application is as follows:

1. A set of visual field test records (310 in total) were used for the purpose of building the classification models. Using the visualized data presented by SOMs and relevant knowledge regarding the visual field test, a domain expert labeled each of these records into either $noisy$: corresponding outliers are measurement noise caused by one of the three behavioral factors: learning, fatigue, and inattention; or $rest$: all other situations. The labeling of these training

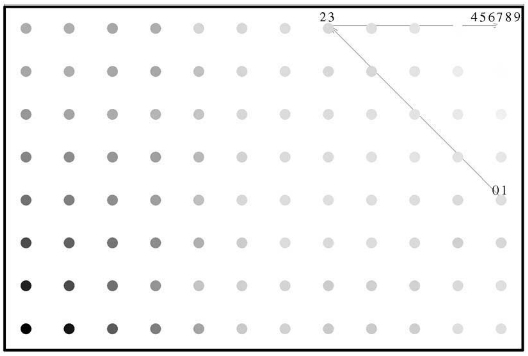

Fig. 5. An example of learning.

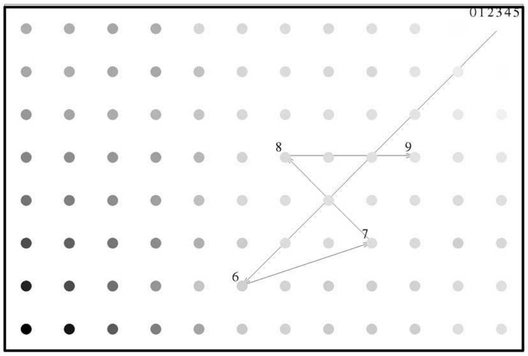

Fig. 6. An example of fatigue.

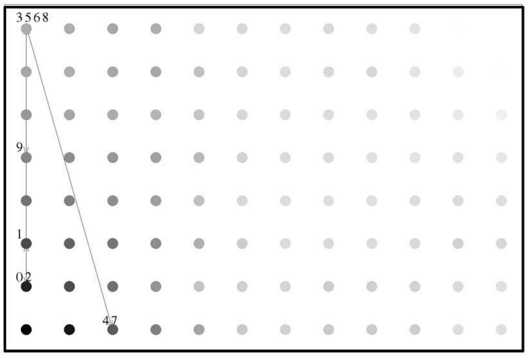

Fig. 7. A glaucoma case (with cluster).

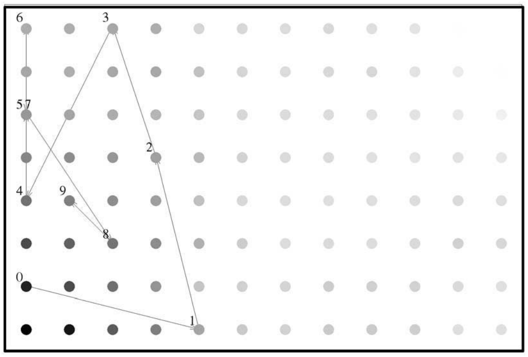

Fig. 8. A glaucoma case (without cluster).

cases into the two classes were made much easier with the help of corresponding SOMs. For example, there is little difficulty in labeling those outliers in Figs. 3, 5, and 6 into *noisy*. When in doubt—when it is not sure whether the outliers in a test record are due to measurement noise or due to pathological factors or when there is no outlier in the data set—put the record into the *rest* class, e.g., the cases illustrated in Figs. 7 and 8.

2. Several features, including the average *sensitivity* (the proportion of positive responses) and the test *stability* (a measure of variations between repeated measurement cycles in a test), were extracted from the data sets and relevant domain knowledge. These features, together with those labeled instances as discussed in the above step, were used to develop the classification models. In particular, we have used Quinlan's C4.5 [18] to learn a set of production rules. Experiments were performed to find a set of rules which would minimize the errors on the unseen cases. This includes the division of 310 cases into training and testing cases of various sizes and the use of a more robust method of 10-fold cross validation. It appears that the 10-fold cross validation presents the most promising results for our data set.
3. Those production rules within the *noisy* class now become our "noise model" and can then be used to delete the corresponding outliers for future test data. Since the noise model is built using those test records with "obvious" noisy outliers, this results in a *cautious* approach to outlier analysis in that only those outliers most likely to be noisy are deleted. Naturally, the confidence of the decisions depends on the quality and coverage of the training examples provided by the expert. Remarkably, with the labeling of just over 300 cases, useful rules have been learned for deleting "noisy" outliers and also for keeping outliers when it is not sure whether they are noisy or not.

Here is one of the learned rules that has been found by the domain expert particularly helpful in identifying the noisy outliers:

Given a test T,
If $stability(T) > 0.60,$
$stability(T) \leq 0.90,$
$average\text{-}sensitivity(T) > 0.77$
Then *the corresponding outliers are noisy.*

4.3 Evaluation

In this section, we present results of evaluating the noise model constructed in the previous section using clinical test data. The visual field test was conducted in a large urban general practice in North London for a glaucoma case findings study. All patients aged 40 years or older who routinely attended the practice for a three-month period during the pilot study were offered the test. A total of 925 patients were screened and 78 of them were later assessed clinically in the practice by an ophthalmologist; this sample included all the 33 people who failed the test and a randomly sampled age matched control group. Among these, 22 eyes were later assessed as glaucoma, 81 were confirmed as normal eyes without any disease, and the rest were diagnosed as other types of ocular abnormalities.

Fig. 9 summarizes the results of examining the *discriminating power* of the test in terms of its glaucoma detection rate versus false alarms using three different data sets: the original 103 test records

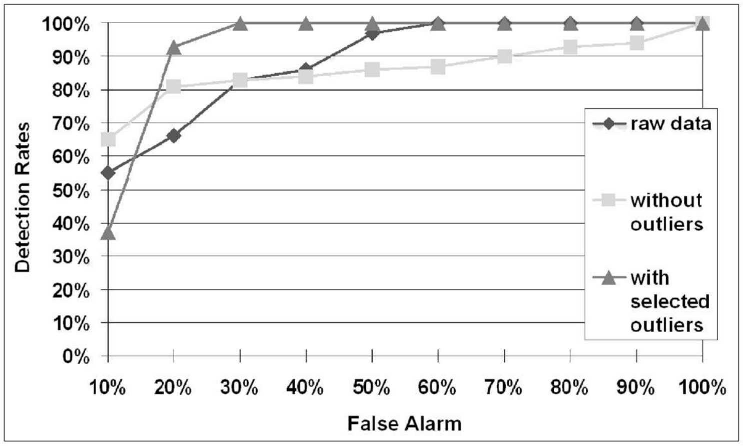

Fig. 9. The results.

corresponding to all the glaucoma and normal eyes, the data set obtained after all outliers are deleted from those test records, and the data set with selected outliers (after applying the noise model to eliminate noisy outliers). The Receive Operator Characteristic (ROC) curves [9] are used to assess the test's diagnostic performance by displaying pairs of false alarms and detection rates throughout the whole range of the test's measurements. While the curves shifted toward the upper left of the diagram, performance of the test is improved in the sense of maximizing detection rates and minimizing false alarms. The decision threshold used for discriminating between normal and glaucoma eyes is the average percentage of positive responses within the test. For instance, the cut-off threshold value of 70 percent has been found to enable the data with selected outliers to achieve a detection rate of 90 percent and a false alarm of 20 percent.

From Fig. 9, it is clear that the data with selected outliers perform better than the other two in terms of maximizing the detection rate and minimizing false alarms. For example, this group can achieve a 100 percent detection rate, while the corresponding false alarm rate is 30 percent. This is equivalent to saying that none of the subjects suffering from glaucoma would have escaped notice and only 30 percent of those normal subjects would have been unnecessarily referred for further examination. To reach a 100 percent detection rate by using the raw data, however, 60 percent of normal subjects would receive false alarms. In comparison with the data with selected outliers, this doubles the number of people who will be referred and further examined unnecessarily.

5 Concluding Remarks

This paper represents a novel attempt in automating the use of domain knowledge in helping distinguish between noisy outliers and *noise-free* outliers, an important issue in many applications including fraud detection, medical tests, process analysis, and scientific discovery. In particular, we have presented a cautious approach to outlier analysis in that only those outliers most likely to be noisy (judged by domain knowledge) are eliminated. This approach to knowledge-based outlier analysis is a useful extension to existing work in both statistical and computing communities on outlier detection. Our research on outlier analysis was initially motivated by a challenging medical application. However, the proposed approach is sufficiently general enough that it may be applied to other applications where the automation of outlier analysis could lead to important benefits. In this note, we have found that AI modeling techniques, when properly integrated, have great potential in automating the challenging knowledge-based outlier analysis process.

Acknowledgments

The work reported in this paper is in part supported by the British Council for Prevention of Blindness, International Glaucoma Association, and The Royal Society, London. An early version of this paper was presented at the Second International Symposium on Intelligent Data Analysis held in London in 1997. The authors would like to thank Sylvie Jami, Phil Dockings, and George Loizou for their general assistance and the IEEE referees for their constructive and helpful comments.

References

[1] V. Barnet, "The Ordering of Multivariate Data (with Discussion)," *J. Royal Statististical Society A,* vol. 139, pp. 318-54, 1976.

[2] V. Barnet and T. Lewis, *Outliers in Statistical Data.* Wiley, 1994.

[3] C. Brodley and M. Friedl, "Identifying and Eliminating Mislabeled Training Instances," *Proc. 13th Nat'l Conf. Artificial Intelligence (AAAI-96),* pp. 799-805, 1996.

[4] K. Carling, "Resistant Outlier Rules and the Non-Gaussian Case," *Computational Statistics and Data Analysis,* vol. 33, no. 3, pp. 249-258, 2000.

[5] P.R. Cohen, *Empirical Methods for Artificial Intelligence.* MIT Press, 1995.

[6] D. Collet and T. Lewis, "The Subjective Nature of Outlier Rejection Procedures," *Applied Statistics,* vol. 25, pp. 228-237, 1976.

[7] R. Gnanadesikan and J.R. Kettenring, "Robust Estimates, Residuals and Outlier Detection with Multi-Response Data," *Biometrics,* vol. 28, pp. 81-124, 1972.

[8] D.J. Hand, *Construction and Assessment of Classification Rules.* Wiley, 1997.

[9] J. Hanely and B. McNeil, "The Meaning and Use of the Area under a Receiver Operator Curve," *Radiology,* vol. 143, pp. 29-36, 1982.

[10] D.M. Hawkins, *Identification of Outliers.* London: Chapman and Hall, 1980.

[11] P.J. Huber, *Robust Statistics.* Wiley, 1981.

[12] B. Kleiner and J. Hartigan, "Representing Points in Many Dimensions by Trees and Castles (with Discussion)," *J. Am. Statistical Assoc.,* vol. 76, pp. 260-276, 1981.

[13] E. Knorr and R. Ng, "A Unified Notion of Outliers: Properties and Computation," *Proc. Third Int'l Conf. Knowledge Discovery and Data Mining (KDD-97),* pp. 219-222, 1997.

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

[15] X. Liu, G. Cheng, and J.X. Wu, "Identifying the Measurement Noise in Glaucomatous Testing: An Artificial Neural Network Approach," *Artificial Intelligence in Medicine,* vol. 6, pp. 401-416, 1994.

[16] X. Liu, G. Cheng, and J.X. Wu, "Noise and Uncertainty Management in Intelligent Data Modeling," *Proc. 12th Nat'l Conf. Artificial Intelligence (AAAI-94),* pp. 263-268, 1994.

[17] N. Matic, I. Guyon, L. Bottou, J. Denker, and V. Vapnik, "Computer Aided Cleaning of Large Databases for Character Recognition," *Proc. 11th Int'l Conf. Pattern Recognition,* pp. 330-333, 1992.

[18] J.R. Quinlan, *C4.5: Programs for Machine Learning.* Morgan Kaufmann, 1993.

[19] S.M. Weiss and C.A. Kulikowski, *Computer Systems that Learn.* Morgan Kaufmann, 1995.

[20] J.X. Wu, "Visual Screening for Blinding Diseases in the Community Using Computer Controlled Video Perimetry," PhD thesis, Univ. of London, 1993.

▷ **For more information on this or any computing topic, please visit our Digital Library at** http://computer.org/publications/dlib.