

# Automated Detection of Outliers in Real-World Data

Mark Last

Department of Information Systems Engineering  
Ben-Gurion University of the Negev  
Beer-Sheva 84105, Israel  
E-mail: mlast@bgumail.bgu.ac.il

Abraham Kandel

Department of Computer Science and Engineering  
University of South Florida  
4202 E. Fowler Avenue, ENB 118  
Tampa, FL 33620, USA  
E-mail: kandel@csee.usf.edu

**Abstract:** Most real-world databases include a certain amount of exceptional values, generally termed as “outliers”. The isolation of outliers is important both for improving the quality of original data and for reducing the impact of outlying values in the process of knowledge discovery in databases. Most existing methods of outlier detection are based on manual inspection of graphically represented data. In this paper, we present a new approach to automating the process of detecting and isolating outliers. The process is based on modeling the human perception of exceptional values by using the fuzzy set theory. Separate procedures are developed for detecting outliers in discrete and continuous univariate data. The outlier detection procedures are demonstrated on several standard datasets of varying data quality.

**Keywords:** Outlier detection, data preparation, data quality, data mining, knowledge discovery in databases, fuzzy set theory.

## 1. Introduction

The statistical definition of an “outlier” depends on the underlying distribution of the variable in question. Thus, Mendenhall et al. [9] apply the term “outliers” to values “that lie very far from the middle of the distribution in either direction”. This intuitive definition is certainly limited to continuously valued variables having a smooth function of probability density. However, the numeric distance is not the only consideration in detecting continuous outliers. The importance of outlier frequency is emphasized in a slightly different definition, provided by Pyle [12]: “An *outlier* is a single, or very low frequency, occurrence of the value of a variable that is far away from the bulk of the values of the variable”. The frequency of occurrence should be an important criterion for detecting outliers in categorical (nominal) data, which is quite common in the real-world databases. A more general definition of an outlier is given in [1]: *an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.*

The real cause of outlier occurrence is usually unknown to data users and/or analysts. Sometimes, this is a flawed value, resulting from the poor quality of a data set, i.e., a data entry or a data conversion error. Physical measurements, especially when performed with malfunctioning

equipment, may produce a certain amount of distorted values. In these cases, no useful information is conveyed by the outlier value. However, it is also possible that an outlier represents correct, though exceptional, information [9]. For example, if clusters of outliers result from fluctuations in behavior of a controlled process, their values are important for process monitoring. According to the approach suggested by [12], recorded measurements should be considered correct, unless shown as definite errors.

### 1.1. Why outliers should be isolated?

The main reason for isolating outliers is associated with data quality assurance. The exceptional values are more likely to be incorrect. According to the definition, given by Wand and Wang [14], unreliable data represents an unconformity between the state of the database and the state of the real world. For a variety of database applications, the amount of erroneous data may reach ten percent and even more [15]. Thus, removing or replacing outliers can improve the quality of stored data. Isolating outliers may also have a positive impact on the results of data analysis and data mining. Simple statistical estimates, like sample mean and standard deviation can be significantly biased by individual outliers that are far away from the middle of the distribution. In

regression models, the outliers can affect the estimated correlation coefficient [10]. Presence of outliers in training and testing data can bring about several difficulties for methods of decision-tree learning, described by Mitchell in [11]. For example, using an outlying value of a predicting nominal attribute can unnecessarily increase the number of decision tree branches associated with that attribute. In turn, this will lead to inaccurate calculation of attribute selection criterion (e.g., information gain). Consequently, the predicting accuracy of the resulting decision tree may be decreased. As emphasized in [12], isolating outliers is an important step in preparing a data set for any kind of data analysis.

## 1.2. Outlier detection and treatment

Manual inspection of scatter plots is the most common approach to outlier detection [10], [12]. Making an analogy with unsupervised and supervised methods of machine learning [11], two types of detection methods can be distinguished: *univariate* methods, which examine each variable individually, and *multivariate* methods, which take into account associations between variables in the same dataset. In [12], a univariate method of detecting outliers is described. According to the approach of [12], a value is considered outlier, if it is far away from other values of the same attribute. However, some very definite outliers can be detected *only* by examining the values of other attributes. An example is given in [10], where one data point stands clearly apart from a bivariate relationship formed by the other points. Such an outlier can be detected only by a multivariate method, since it is based on dependency between two variables. Manual detection of outliers suffers from the two basic limitations of data visualization methods: subjectiveness and poor scalability (see [6]). The analysts have to apply their own subjective perception to determine the parameters like “very far away” and “low frequency”. Manual inspection of scatter plots for every variable is also an extremely time-consuming task, not suitable for most commercial databases, containing hundreds of numeric and nominal attributes.

An objective, quantitative approach to unsupervised detection of numeric outliers is described in [9]. It is based on the graphical technique of constructing a *box plot*, which represents the *median* of all the observations and two *hinges*, or medians of each half of the data

set. Most values are expected in the *interquartile range* ( $H$ ) located between the two hinges. Values lying outside the  $\pm 1.5H$  range are termed “mild outliers” and values outside the boundaries of  $\pm 3H$  are termed “extreme outliers”. While this method represents a practical alternative to manual inspection of each box plot, it can deal only with continuous variables characterized by unimodal probability distributions. The other limitation is imposed by the ternary classification of all values into “extreme outliers”, “mild outliers”, and “non-outliers”. The classification changes abruptly with moving a value across one of the  $1.5H$  or  $3H$  boundaries.

An information-theoretic approach to supervised detection of erroneous data has been developed by Guyon *et al.* in [4]. The method requires building a prediction model by using one of data mining techniques (e.g., neural networks or decision trees). The most “surprising” patterns (having the lowest probability to be predicted correctly by the model) are suspicious to be unreliable and should be treated as outliers. However, this approach ignores the fact that data conformity may also depend on the inherent distribution of database attributes and some subjective, user-dependent factors.

Zadeh [16] applies the fuzzy set theory to calculating the *usuality* of given patterns. The notion of usuality is closely related to the concept of *disposition* (proposition which is preponderantly but not necessarily true). In [16], the fuzzy quantifier *usually* is represented by a fuzzy number of the same form as *most*. One example of a disposition is *usually it takes about one hour to drive from A to B*. The fuzzy set of normal (or regular) values is considered the complement of a set of *exceptions*. In the above example, a five-hour trip from A to B would have a higher membership grade in the set of exceptions than in the set of normal values. In [8], we have presented an advanced method, based on the information theory and fuzzy logic, for measuring reliability of multivariate data. The fuzzy degree of reliability depends on two factors:

- 1) The distance between the value predicted by a data mining model and the actual value of an attribute.
- 2) The user perception of “unexpected” data.

Rather than partitioning the attribute values into two “crisp” sets of “outliers” and “non-outliers”, the fuzzy logic approach assigns a continuous degree of reliability (ranging between 0.0 and

1.0) to each value. In this paper, we are extending the approach of [8] to detection of outliers in univariate data.

No matter how the outliers are detected, they should be handled before applying a data mining procedure to the data set. One trivial approach is to discard an entire record, if it includes at least one outlying value. This is similar to ignoring records containing missing values by some data mining methods. An alternative is to correct (“rectify”) an outlier to another value. Thus, Pyle [12] suggests a procedure for remapping the variable’s outlying value to the range of valid values. Outliers detected by a supervised method, based on the linear regression model, can be adjusted iteratively by assigning to each observation a weight depending on its residual from the fitted value [3]. Actually, if the outlying value is assumed completely erroneous, the correct value can be estimated by any method for estimating missing attribute values (see [11]).

### 1.3. Paper organization

This paper is organized as follows.

In Section 2, we present novel, fuzzy-based methods for univariate detection of outliers in discrete and continuous attributes. In Section 3, the methods are applied to several machine learning datasets of varying size and quality. Section 4 concludes the paper with some directions for future research in outlier detection.

## 2. Fuzzy-based detection of outliers in univariate data

### 2.1. Detecting outliers in discrete variables

According to [9], a discrete variable is assumed to have a *countable* number of values. On the other hand, a continuous variable can have *infinitely many values* corresponding to the points on a line interval. In practice, however, the distinction between these two types of attributes may be not so clear (see [12]). The same attributes may be considered discrete or continuous, depending on the precision of measurement, and other application-related factors. For the purpose of our discussion here, we assume the discrete attributes to include binary-valued (dichotomous) variables,

categorical variables, and numeric variables with a limited number of values.

Actually, any single or low frequency occurrence of a value may be considered an outlier.

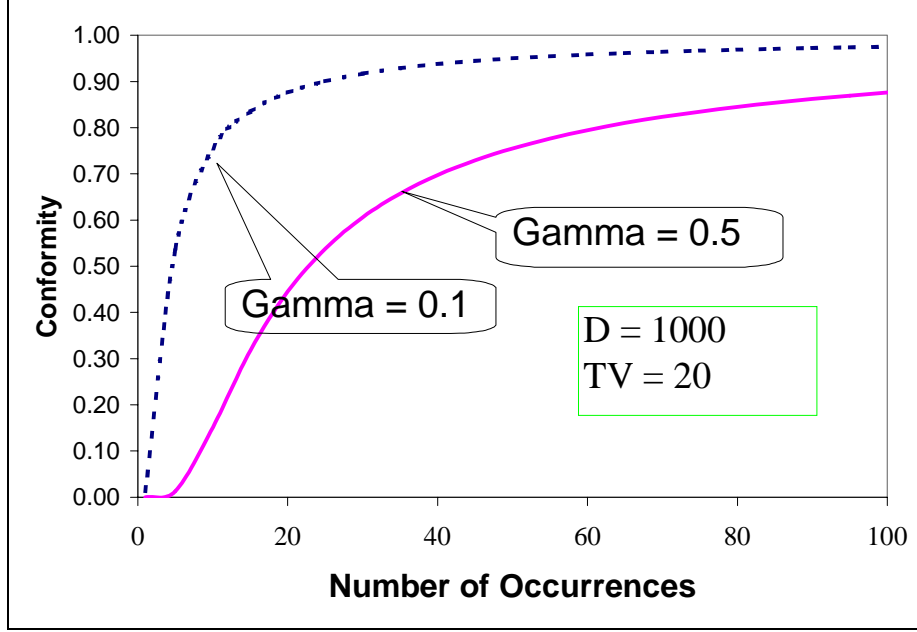
However the human perception of an outlying (rare) discrete value depends on some additional factors, like the number of records where the value occurs, the number of records with a non-null value, and the total number of distinct values in an attribute. For instance, a single occurrence of a value in one record out of twenty does not seem very exceptional, but it is an obvious outlier in a database of 20,000 records. The value taken by only 1% of records is clearly exceptional if the attribute is binary-valued, but it is not necessarily an outlier in an attribute having more than 100 distinct values. Our attitude to rare values also depends on the objectives of our analysis. For example, in a marketing survey, the records of potential buyers represent a very important part of the population, even if they constitute just one percent of the entire sample.

To automate the cognitive process of detecting discrete outliers, we represent the *conformity* of an attribute value by the following membership function  $\mu_R$ :

$$\mu_R(V_{ij}) = \frac{2}{1 + e^{\frac{D_i}{\gamma \cdot TV_i \cdot N_{ij}}}} \quad (1)$$

Where  $V_{ij}$  is the discrete value No.  $j$  of the attribute  $A_i$ ;  $\gamma$  is the shape factor, representing the user attitude towards conformity of rare values;  $D_i$  is the total number of records where  $A_i \neq null$ ;  $TV_i$  is the total number of distinct values taken by the attribute  $A_i$ ; and  $N_{ij}$  is the number of occurrences of the value  $V_{ij}$ .

It can be easily verified that the Equation (1) agrees with the definition of *fuzzy measure* (see [5]). The conformity becomes close to zero, when the number of value occurrences is much smaller than the average number of records per value (given by  $D_i / TV_i$ ). On the other hand, if the data set is very small ( $D_i$  close to zero),  $\mu_R$  approaches one, which means that even a single occurrence of a value is not considered an outlier.



**Figure 1: Discrete value conformity as a function of the shape factor  $\gamma$**

The shape factor  $\gamma$  represents the subjective attitude of a particular data analyst to values of the same frequency. Low values of  $\gamma$  (about 0.5) make  $\mu_R$  a sigmoidal function gradually increasing from zero to one with the number of occurrences  $N_{ij}$ . For  $\gamma = 0$ , all values are considered conforming ( $\mu_R = 1$ ). On the other hand, high values of  $\gamma$  (more than one) make  $\mu_R$  a step function, marking almost any value as “outlier” ( $\mu_R = 0$ ). In Figure 1, we show  $\mu_R$  as a function of  $N_{ij}$  for two values of the factor  $\gamma$ . For our case studies in the next Section, we have used the value of  $\gamma = 0.5$ .

The subset of outliers in each attribute is found by defining an  $\alpha$ -cut of all the attribute values:  $\{V_{ij} : \mu_R(V_{ij}) < \alpha\}$ . In our applications (see Section 3 below), we use  $\alpha = 0.05$ . If an attribute is dichotomous (has only two distinct values) and one of the values is recognized as an outlier, we can remove that attribute from the subsequent knowledge discovery process and even from the original database, since it does not provide us with any new information about specific records. Thus, in some cases, the outlier detection can be used for *dimensionality reduction* of data.

## 2.2. Detecting outliers in continuous variables

The sets of discrete and continuous data are not completely disjoint. Numeric attributes taking a small number of values can be treated as discrete variables, and categorical values can be translated into consecutive numbers (examples of such transformations are represented in [12]). However, the criterion of value *frequency* (see sub-section 2.1 above) cannot be applied directly to detecting continuous outliers, since a continuous attribute can take an infinite number of values in its range. In the continuous case, we should *sort* the distinct values of the inspected attribute and examine the *distance* between each single value and its neighboring values. Greater that distance, lower is our confidence in a value (e.g., see the rightmost point in Figure 2a). Still we may have several clusters of attribute values, located at a large distance from each other (see Figure 2b). None of these clusters may contain outliers.

To automate the human perception of non-conforming (outlying) continuous values, we need a formal definition of value conformity, with respect to its preceding and succeeding values. Such a definition is suggested below.

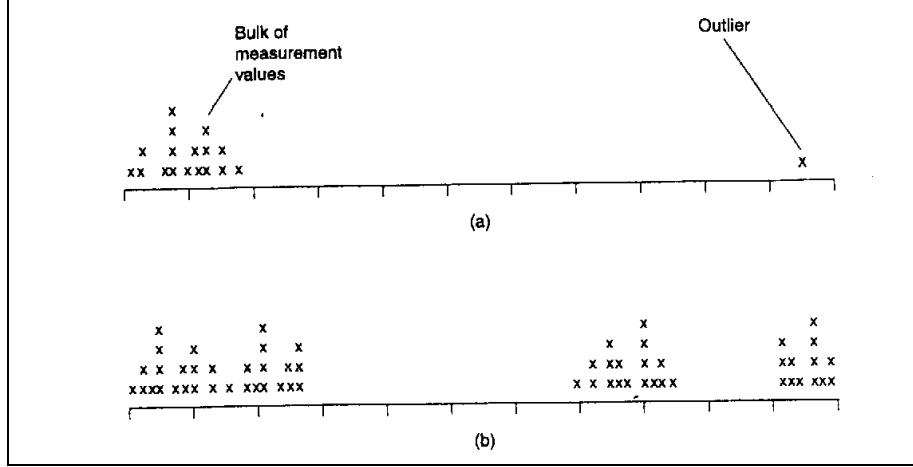


Figure 2: Examples of outliers (from[12])

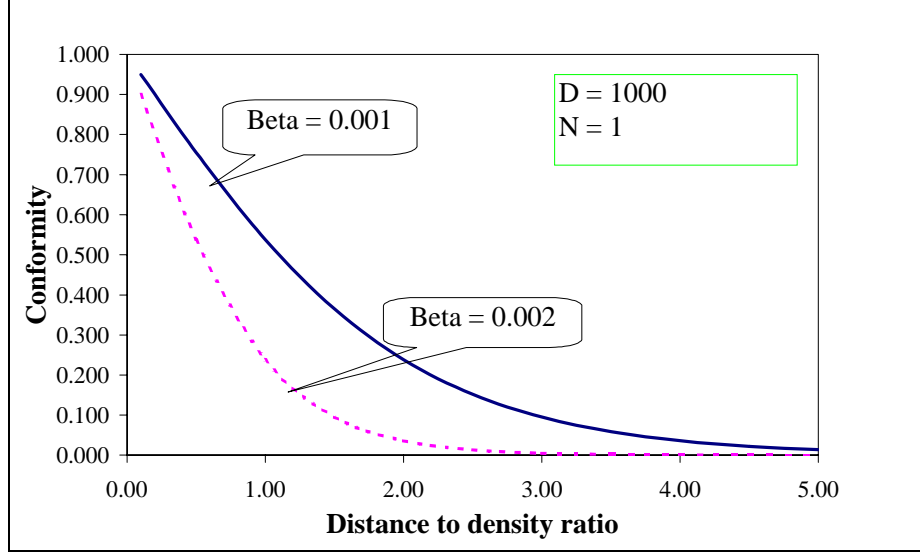
**Definition.** Given a set of sorted distinct values, a value is considered *conforming from below* (*conforming from above*), if it is close enough to the succeeding (preceding) values respectively. Thus, automated detection of outliers requires sorting the continuous attribute values in ascending order (from  $V_{i1}$  to  $V_{i,TV_i}$ , where  $TV_i$  is the total number of distinct values). Afterwards, the conformity of each continuous value is measured with respect to its preceding and succeeding values. The conformity from below is denoted by  $\mu_{RL}$  and the conformity from above is denoted by  $\mu_{RH}$ . Each conformity measure depends on the so-called “distance to density” ratio. For the conformity from below, this is the ratio between the distance from each value to the subsequent value and the average density of  $M$  subsequent values ( $M$  is termed the “look-ahead”). Accordingly, the conformity from above is determined by the ratio between the distance from each value to the preceding value and the average density of  $M$  preceding values. Other factors involved in the conformity calculation include the shape factor  $\beta$ , the total number of records  $D_i$ , and  $N_{ij}$ , the number of

occurrences of the value  $V_{ij}$ . The expressions for the membership functions  $\mu_{RL}$  and  $\mu_{RH}$  follow.

$$\mu_{RL} = \frac{2}{1 + e^{\frac{\beta \cdot M \cdot D_i \cdot (V_{i,j+1} - V_{ij})}{N_{ij} \cdot (V_{i,j+M+1} - V_{i,j+1})}}} \quad (2)$$

$$\mu_{RH} = \frac{2}{1 + e^{\frac{\beta \cdot M \cdot D_i \cdot (V_{ij} - V_{i,j-1})}{N_{ij} \cdot (V_{i,j-1} - V_{i,j-M-1})}}} \quad (3)$$

The shape factor  $\beta$  represents the user-dependent attitude to distances between succeeding values. Lower values of  $\beta$  (about  $10^{-4}$ ) assign conformity of 0.5 even to a single value, having a 10 times larger distance to the neighboring value than the average density of succeeding (or preceding) values. Higher values of  $\beta$  (like  $10^{-3}$ ) provide a sharper decrease of conformity from 1.0 to zero, as the distance between succeeding values becomes larger. In Figure 3, we show  $\mu_{R*}$  as a function of the “distance to density” ratio for two different values of  $\beta$  (0.001 and 0.002), given that the value occurs only in one record out of 1000. The value of  $\beta = 0.001$  has been used for in our case studies (see Section 3).



**Figure 3: Conformity of a continuous value as a function of the shape factor  $\beta$**

The procedure of determining the conformity of each continuous value depends on the available prior knowledge about the distribution of all the attribute values. If our prior knowledge causes us to assume that the distribution is *unimodal* (making the situation in Fig. 2b unlikely), the outlier detection can be focused on checking only the conformity “from below” of the lowest values and the conformity “from above” of the highest values. Moreover, we can stop the search for outliers, once we have found the first conforming value from each side because a unimodal distribution is supposed to have only one cluster of values. The pseudocode of outlier detection in unimodal attributes is given in the algorithm A below.

**Algorithm A. Outlier detection in Unimodal Attributes**

- Select the look-ahead  $M$ , the threshold  $\alpha$ , and the shape factor  $\beta$
  - Sort distinct values in ascending order
  - Initialize conformity of each value to 1.00 ( $\forall j: \mu_{RL}(V_{ij}) = \mu_{RH}(V_{ij}) = 1$ )
- Calculating conformity from below*
- Initialize the index of current value to zero (index of the lowest value)
  - Do
    - Calculate the “low value” conformity  $\mu_{RL}(V_{ij})$  for the current value by Equation (2)
    - Increment the index of the current value by one

- While the index of current value  $< (Total\_Number\_of\_Values - M - 2)$  and  $\mu_{RL}(V_{ij}) < \alpha$
  - Set the lower bound of the attribute (*Low\_Bound*) to the current value  $V_{ij}$
- Calculating conformity from above*
- Initialize the index of current value to  $Total\_Number\_of\_Values - 1$  (index of the highest value)
  - Do
    - Calculate the “high value” conformity  $\mu_{RH}(V_{ij})$  for the current value by Equation (3)
    - Decrement the index of the current value by one
  - While the index of current value  $> M + 1$  and  $\mu_{RH}(V_{ij}) < \alpha$
  - Set the upper bound of the attribute (*Upper\_Bound*) to the current value  $V_{ij}$
  - *Finding outlying values*
  - For each value  $V_{ij}$ ,
    - If  $(V_{ij} < Low\_Bound)$  or  $(V_{ij} > Upper\_Bound)$ 
      - Denote  $V_{ij}$  as outlier
  - End
- When no information about the attribute distribution is available, we can consider as an outlier only a value, which is far away from both its preceding and succeeding values. That is, in this case we should calculate two conformity degrees for each value:  $\mu_{RL}$  and  $\mu_{RH}$ . An outlying value should have *both* conformity degrees below the threshold  $\alpha$ . If only one conformity degree is below  $\alpha$ , this means that a value is the highest or the lowest in a cluster, but not

necessarily an outlier. The algorithm for detecting outliers in continuous attributes of unknown distribution follows.

**Algorithm B.** *Outlier detection in Attributes of Unknown Distribution*

- Select the look-ahead  $M$ , the threshold  $\alpha$ , and the shape factor  $\beta$
- Sort distinct values in ascending order
- Initialize reliabilities of each value to zero ( $\forall j: \mu_{RL}(V_{ij}) = \mu_{RH}(V_{ij}) = 0$ )

*Calculating conformity from below*

- For  $index\_of\_current\_value = 0$  to  $index\_of\_current\_value = (Total\_Number\_of\_Values - M - 3)$ 
  - Calculate the “low value” conformity  $\mu_{RL}(V_{ij})$  for the current value by Equation (2)

*Calculating conformity from above*

- For  $index\_of\_current\_value = (Total\_Number\_of\_Values - 1)$  to  $index\_of\_current\_value = M + 2$ 
  - Calculate the “high value” conformity  $\mu_{RH}(V_{ij})$  for the current value by Equation (3)
- *Finding outlying values*
- For each value  $V_{ij}$ ,
  - If  $\max\{\mu_{RL}, \mu_{RH}\} < \alpha$ 
    - Denote  $V_{ij}$  as outlier
- End

### 3. Evaluating quality of machine learning datasets

We have applied the automated methods of detecting outliers in discrete and continuous attributes (see Section 2 above) to seven datasets from the UCI Machine Learning Repository [2]. The datasets in this repository are widely used by the data mining community for the empirical evaluation of learning algorithms. There are many studies comparing the performance of different classification methods on these datasets, but we do not know about any documented attempt of evaluating their quality. Though these datasets are supposed to represent the “real-world” data, their documentation sometimes mentions certain amount of manual cleaning, performed by the data providers. The datasets selected by us for the outlier detection contain a diverse mixture of discrete and continuous attributes. The brief description of each dataset follows.

*The Breast Cancer Database.* This is a medical data set including 699 clinical cases. There are

nine discrete multi-valued attributes (ranging from 1 to 10) that represent results of medical tests and one discrete binary-valued attribute standing for the class of each case. The documentation indicates that several records have been removed from the original dataset (probably, due to data quality problems). *Chess Endgames.* This is an artificial data set representing a situation in the end of a chess game. Each of 3,196 instances is a board-description for this chess endgame. There are 36 attributes describing the board. All the attributes are nominal (mostly, binary). Since the data was artificially created, no data cleaning was performed.

*Credit Approval.* This is an encoded form of a proprietary database, containing data on credit card applications for 690 customers. The data set includes eight discrete attributes (with different number of distinct values) and six continuous attributes. There were originally a few missing values, but these have all been replaced by the overall median. There is no information about any other cleaning operations applied.

*Diabetes.* This is a part of larger diagnostic dataset, and it contains data on 768 patients. In the dataset, there are eight continuous attributes and one binary-valued attribute. No data cleaning is mentioned in the dataset documentation.

*Glass Identification.* This database deals with classification of types of glass, based on chemical and physical tests. It includes 214 cases, with nine continuous attributes and one discrete attribute having seven distinct values. Using any kind of data cleaning by the data donors is not mentioned.

*Heart Disease.* The dataset includes results of medical tests aimed at detecting a heart disease for 297 patients. There are seven nominal (including three binary-valued) and six continuous attributes. Other 62 attributes presenting in the original database have been excluded from the dataset. According to the documentation, missing values were encoded with a pre-defined value.

*Iris Plants.* R.A. Fisher has created this database in 1936 and, since then, it has been extensively used in the pattern recognition literature. The data set contains three classes of 50 instances each, where each class refers to a type of iris plant. Each instance is described by four continuous attributes.

The automated methods of outlier detection have been applied to the above datasets by using the

following parameters (tuned to the perception of the authors on small samples of graphically represented data):

- Minimum conformity of non-outliers ( $\alpha$ ): 0.05
- Shape factor for calculating conformity of continuous values ( $\beta$ ): 0.001
- Shape factor for calculating conformity of discrete values ( $\gamma$ ): 0.500
- Look-ahead for calculating average density of consecutive values ( $M$ ): 10

Algorithm A (see above) has been applied to all continuous attributes, assuming their distribution to be unimodal. This assumption was verified by comparing the outputs of Algorithm A and Algorithm B. The results of outlier detection in all datasets are represented in Table 1 below. In the third and the fourth columns from the left, we can see the number and the percentage of “non-conforming records”, containing at least one

outlying value. The Iris dataset is the most “clean” one: it contains no outliers at all. This finding is not surprising, taking into account the popularity of Iris in the Machine Learning community. Most other datasets, tested by us, have a small amount of outliers in their records (between 0.5% in the Glass and 8.3% in the Breast). However, the Chess dataset suffers from a significant portion of outlying values – more than 25% of its records contain at least one outlier! The outliers are found in 12 binary-valued attributes, where one of the values has a very low frequency (e.g., in *spcop* attribute, there is one record having the value of 0 vs. 3,195 records having the value of 1). These attributes can be completely discarded from the data mining process. As indicated above, our method of automated outlier detection can also be used as a dimensionality reduction tool.

Dataset	Total records	Non-conforming Records	Percentage	Discrete Attributes		Continuous Attributes	
				Total	Containing outliers	Total	Containing outliers
Breast	699	58	8.3%	10	6	0	0
Chess	3196	838	26.2%	37	12	0	0
Credit	690	41	5.9%	9	4	6	5
Diabetes	768	14	1.8%	1	0	8	6
Glass	214	1	0.5%	1	0	9	1
Heart	297	5	1.7%	8	1	6	1
Iris	150	0	0.0%	1	0	4	0

**Table 1.** Detecting outliers in the Machine Learning Datasets

In columns no. 5 – 8 of Table 1, we summarize the conformity of discrete and continuous attributes in each dataset. One example is the popular Credit Approval, comprising a mixture of discrete and continuous attributes. Since this dataset is based on real-world banking data, it is quite reasonable that four discrete attributes (out of nine) and five continuous attributes (out of six) contain outlying values. In Figure 4, we show the number of occurrences (solid bars) and the calculated conformity (a solid line) for each discrete value of the attribute “Job Status” (denoted as A6 in the encoded versions of the

Credit Approval dataset). Visually, the values 2, 3, 7, and 9 seem to be clear outliers, confirming the results of the automated outlier detection (their conformity is below the threshold  $\alpha$ ). demonstrates detection of outliers in a continuous attribute (“Age”, or A2). Only the lower values of this attribute are shown (18 years and younger). When looking at the chart, the 13.75 years old person seems considerably younger than most other customers. The automated method of outlier detection has assigned to this value the conformity “from below” of 0.007, marking it as an outlier.



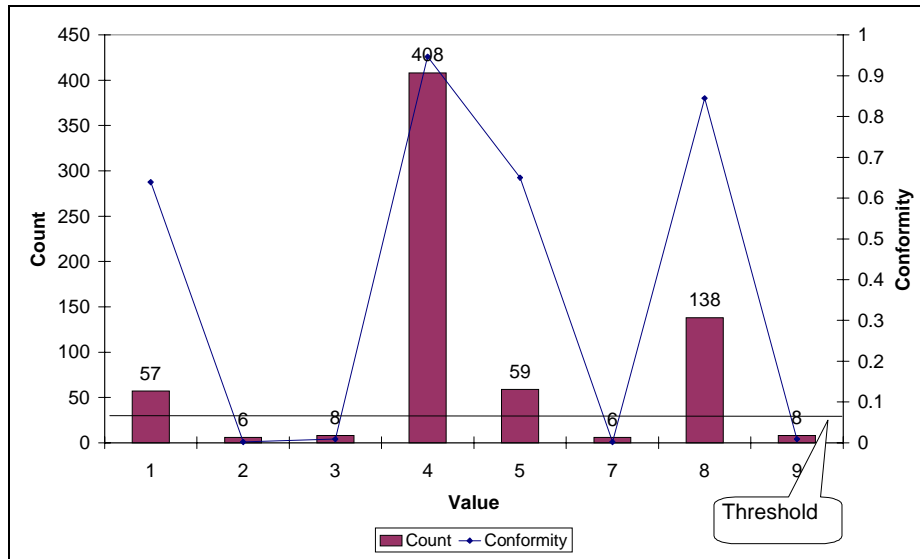


Figure 4: Detecting outliers in a discrete attribute (“Job Status”)

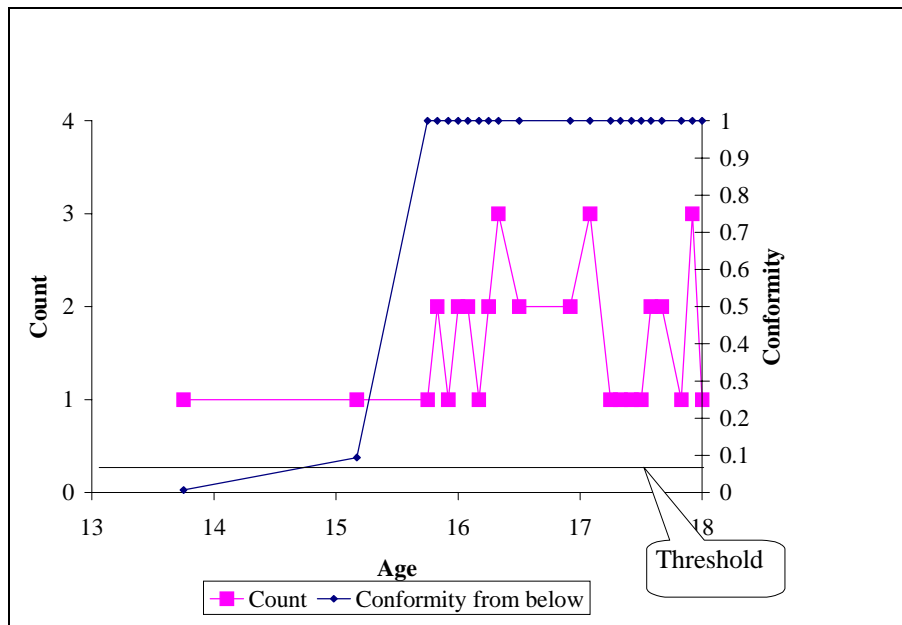


Figure 5: Detecting outliers in a continuous attribute (“Age”)

#### 4. Conclusions

In this paper, we have presented novel, fuzzy-based procedures for automating the human perceptions of outlying values in univariate data. The prior knowledge about the data can be utilized to choose the appropriate procedure for a given dataset, and further to adapt the process of outlier detection to the form of attribute distribution (e.g., unimodal distribution). The high dimensionality of modern databases provides a significant advantage to automated

perceptions of outliers over the manual analysis of visualized data. Unlike the case of human decision-making, the parameters of the automated detection of outliers can be completely controlled, making it an objective tool of data analysis. The results of outlier detection can be used to improve the quality of data in a database and, in some cases, to enhance the performance of data mining algorithms. As already shown by us in [6], the fuzzy set theory enables to develop computational models of human perception for the problems, where the

graphically represented data has been traditionally analyzed by human experts. In this paper, we have described one procedure for detecting outliers in discrete data and two procedures of outlier detection in continuous data. More techniques may be developed by catching different aspects of human perception and making use of prior knowledge, available about attributes, including domain size, validity range and form of distribution. The automated detection of outliers may be embedded in database management systems, to warn the users against possibly inaccurate information. The integration of outlier detection with data mining methods has a potential benefit for extracting valid knowledge from “dirty” data.

*Acknowledgment.* This work was supported in part by the National Institute for Systems Test and Productivity at USF under the USA Space and Naval Warfare Systems Command Grant No. N00039-01-1-2248 and in part by the USF Center for Software Testing under grant No. 2108-004-00.

## References

- [1] Barnett, V., Lewis, T. (1995). *Outliers in Statistical Data*. Wiley, 3<sup>rd</sup> Edition.
- [2] Blake, C., Keogh, E. & Merz, C.J. (1998). *UCI Repository Of Machine Learning Databases* [<http://www.ics.uci.edu/~mllearn/MLRepository.html>]. Irvine, CA: University of California, Department of Information and Computer Science.
- [3] Fan, J., & Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman & Hall.
- [4] Guyon, I., Matic, N., & Vapnik, V. (1996). Discovering Informative Patterns and Data Cleaning. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* pp. 181-203. Menlo Park, CA: AAAI/MIT Press.
- [5] Klir, G. J., & Yuan, B. (1995). *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Upper Saddle River, CA: Prentice-Hall.
- [6] Last, M., & Kandel, A. (1999). Automated Perceptions in Data Mining. *Proceedings of the Eighth International Conference on Fuzzy Systems*. Seoul, Korea. Part I, pp. 190 – 197.
- [7] Maimon, O., & Last M. (2000). *Knowledge Discovery and Data Mining – The Info-Fuzzy Network (IFN) Methodology*. Kluwer Academic Publishers.
- [8] Maimon, O., Kandel, A. & Last, M. (2001). Information-Theoretic Fuzzy Approach to Data Reliability and Data Mining. *Fuzzy Sets and Systems*. Vol. 117, No. 2, pp. 183-194.
- [9] Mendenhall, W., Reinmuth, J.E., & Beaver, R.J. (1993). *Statistics for Management and Economics*. Belmont, CA: Duxbury Press.
- [10] Minium, E.W., Clarke, R.B., & Coladarci, T. (1999). *Elements of Statistical Reasoning*. New York: Wiley.
- [11] Mitchell, T.M. (1997). *Machine Learning*. New York: McGraw-Hill.
- [12] Pyle, D. (1999). *Data Preparation for Data Mining*. San Francisco, CA: Morgan Kaufmann.
- [13] Quinlan, J.R. (1986). Induction of Decision Trees. *Machine Learning*, 1, 1, 81-106.
- [14] Wand, Y., & Wang, R.Y. (1996). Anchoring Data Quality Dimensions in Ontological Foundations. *Communications of the ACM*, 39, 11, 86-95.
- [15] Wang, R.Y., Reddy, M.P., & Kon, H.B. (1995). Toward Quality Data: An Attribute-based Approach. *Decision Support Systems*, 13, 349-372.
- [16] Zadeh, L.A. (1985). Syllogistic Reasoning in Fuzzy Logic and its Application to Usuality and Reasoning with Dispositions. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-15, 6, 754-763.