Intelligent Data Analysis 10 (2006) 521–538
IOS Press

A comprehensive survey of numeric and symbolic outlier mining techniques

Malik Agyemang, Ken Barker and Rada Alhajj
Department of Computer Science, University of Calgary, 2500 University Drive N.W. Calgary, AB, Canada T2N 1N4
E-mail: {agyemang,barker,alhajj}@cpsc.ucalgary.ca

Received 16 August 2005
Revised 11 November 2005
Accepted 17 December 2005

Abstract. Data that appear to have different characteristics than the rest of the population are called outliers. Identifying outliers from huge data repositories is a very complex task called *outlier mining*. Outlier mining has been akin to finding needles in a haystack. However, outlier mining has a number of practical applications in areas such as fraud detection, network intrusion detection, and identification of competitor and emerging business trends in e-commerce. This survey discuses practical applications of outlier mining, and provides a taxonomy for categorizing related mining techniques. A comprehensive review of these techniques with their advantages and disadvantages along with some current research issues are provided.

Keywords: Symbolic, rule-based, distance-based, depth-based, distribution-based, outliers, interestingness, unexpectedness, taxonomy, web-based, exception patterns

1. Introduction

The complexity of today's databases (i.e., volume and dimension) calls for automated tools if their contents are to be efficiently analysed to expose patterns hidden in them for decision support systems. The automated process of finding useful information from data is called data mining. Data mining is described as the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. It involves the entire process of extracting useful information and subsequent analysis of the extracted information for decision making purposes [7,34]. Fayyad et al. [24] defines knowledge discovery in databases as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. There are several data mining techniques, notable among them are association rule mining, classification and prediction, clustering, evolution/time series mining, and outlier/exception mining, etc. [34]. Except outlier mining, the other techniques deal with finding patterns associated with the majority of the data elements. However, exception/outlier mining is dedicated to finding rare patterns associated with very few of the data objects called *outliers*.

Outliers are data that do not obey rules considered normal for the majority of the data elements [20, 27]. Outliers may occur as a result of mechanical faults (e.g., faulty sensors recording incorrect hourly temperature readings), changes in system behavior, fraudulent behavior (e.g., illegally using someone's

credit/debit card) or through natural deviations in population (e.g., very low infant mortality rate for Japanese-American living in the East Coast of the United States [53], prevalence of sickle cell anaemia among Africans, humans measuring more than 8 feet, etc). Identifying these uncommon occurrences *a priori* may help stakeholder in making policy. For example, the discovery of the very low infant mortality rates (i.e., far below the national average) among Japanese-Americans overwhelmed public health practitioners which resulted in more funded research in mining public health data. Similarly, detecting credit card or phone cloning fraud before they happen may save the service providers millions of dollars in lost income. Similarly, identifying network intrusions *a priori* may save companies millions of dollars in lost revenue.

There are two schools of thought about outliers. The first treats outliers as errors or noise that must be removed during preprocessing to ensure accurate data mining process. Techniques based on this principle are either built to be resistant to the effects of outliers (e.g. [63]) or remove them as errors before the analysis begins (e.g., the ROR algorithm by Adam et al. [11] is able to removes about 80% of the raw data if they are identified as real outliers). The second school of taught treats outliers as potentially interesting and surprising patterns hidden in data. The ultimate goal of techniques based on this principle is to find rare and interesting patterns for decision making as oppose to removing them. This paper follows the latter school of though and provides practical applications for outlier mining. Potential applications include:

Fraud Detection: A major problem for credit card companies is illegal use of stolen or lost cards. Detecting and preventing the use of such cards is critical because credit card companies assume liability for unauthorized expenses on stolen credit cards. Chan et al. [23] propose outlier mining techniques that combine multiple fraud detectors to significantly reduce the financial losses to service provider through fraud. Bolton and Hang [15] propose the behavioral outlier detection algorithm that tracks the abnormal spending behavior and the frequency of the transactions involved. Transactions considered abnormal from normal behaviors are identified as outliers. A very good survey of fraud detection techniques is provided by Kou et al. [41].

Sports: Outlier mining algorithms are applied to the NHL-player statistics from 1995 to reveal the most outstanding players based on different attributes. Gretzky and Fedorov are identified as the most outstanding players using *points scored*, *plus-minus statistics*, and *penalty minutes*. Similar interesting results are obtained with the German National Soccer League [42,43].

Health: Outlier mining techniques can be used to determine tumors that will never develop to full cancer. It can also be used to determine specific types of cancer cells that may never respond to treatment and hence become fatal. Other rare patterns hidden in health data could ultimately prove to be very valuable, Provost et al. [53] discovered, from large public health data that Japanese-Americans living in the East Coast of the United States have very low infant mortality rate (0.18%) which is far below the national average. This amazing result was unknown to health authorities.

Politics: He et al. [30] applied a semantic outlier mining algorithm to determine interesting voting patterns in the US senate. Their analysis on 16 issues at the senate floor show a Republican who voted on 8 issues like a Democrat and 5 issues like a Republican. Thus, the senator is more of a Democrat than a Republican with respect to the 16 issues on the senate floor.

E-commerce: The application of outlier mining techniques can lead to the identification of customers who may easily change their minds while surfing the web [36]. It can identify useful information from a competitor's website using the information contained in the web pages [50], and the identification of competitors and emerging business trends [2].

Others: In addition to the applications discussed above, traditional outlier mining techniques have been applied to find outliers in other real-world scenarios such as weather prediction and metrological

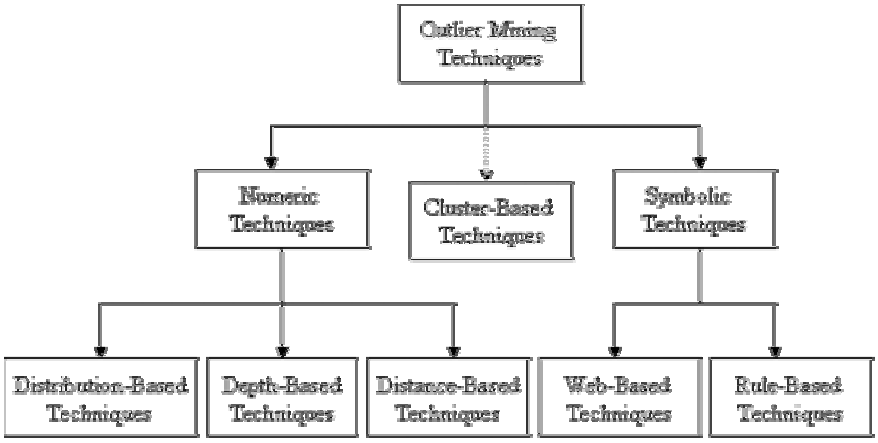

Fig. 1. Taxonomy of Outlier Mining Techniques.

data analysis [79], intrusion detection [41,54], identification of criminal behaviour [47], identification of graph-based spatial outliers [66], time series data analysis [39,77], etc. Hodge et al. [26] outlines several practical applications of outlier mining.

This survey identifies three categories of outlier mining techniques namely, numeric, symbolic, and cluster-based methods. The numeric methods consist of the distribution-based, depth-based, and distance-based outlier mining techniques. The symbolic techniques are divided into web-based and rule-based. The cluster-based techniques are identified as "partial class" because they are not explicitly outlier mining techniques. They are clustering algorithms with a potential of producing outliers as their by-product. The advantages and disadvantages of these techniques are also discussed. The outlier mining taxonomy is shown in Fig. 1 with detailed description of the components given subsequently.

2. Numeric outlier mining techniques

The numeric outlier mining techniques consist of all techniques that use numeric data as direct inputs. These techniques are subdivided into distribution-based, depth-based, and distance-based. The distance-based and the depth-based techniques are mainly from computation statistics and are distinguished in that the former fits data with standard statistical distributions (e.g., normal, Poisson, etc.), while latter does not. On the other hand, the distance-based techniques consist of all numeric techniques that involve computing some form of a metric distance. Some of these techniques may apply clustering algorithms as a means to identify outliers (e.g. [43,62]). Cluster-based techniques are discussed in Section 3.

2.1. The distribution-based techniques

The distribution-based techniques are based on sound mathematical methods from applied statistics and probability theory. They involve fitting data with standard probabilistic models and observing the behaviour of the data in relation to the model. Data objects that conform to the distribution laws of the model are considered normal while those that deviate from these laws are considered outliers. The probabilistic models are necessary but not sufficient for determining outliers. Procedures that determine whether or not a particular point is an outlier in relation to a standard model, called discordancy tests,

are required. The literature on outlier detection reveals over a hundred discordancy tests designed for detecting outliers under different conditions [20,59].

A standard discordancy test consists of verifying the *null hypothesis* against the *alternative* hypothesis. The *null* hypothesis is the basic statement that *data have been generated using a given distribution law* (e.g., normal, gamma, etc.). The *alternative* hypothesis is the opposite of the null hypothesis. If there is enough information to support the *null* hypothesis for a particular record then that record is normal otherwise it is an outlier. Statistically, rejecting the null hypothesis means the data being tested have been generated with a distribution different than the null hypothesis. The choice of an appropriate discordancy test for a given data depends on: knowledge of the data distribution, knowledge of the distribution parameters, and the number and type of expected outliers needed, which accounts for the many discordancy tests available for finding distribution-based outliers [20,26].

The regression model for outlier detection involves finding the dependence of a set of random variable(s) Y on another set of random variable(s) X. The data containing the outliers is fitted with a standard regression model and point(s) with the greatest error is/are removed for further testing [21,28]. Torr and Murray [74] used regression analysis as a diagnostic outlier detector. They fit the data points with a least square regression to minimise Eq. (1):

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2 \tag{1}$$

where $\hat{Y}$ is the mean of the data. The point with maximal influence on the regression line is removed and the remaining data points refitted with the regression line. The process continues until there are no more outliers in the dataset. The set of data removed constitute the outlier set. To make the least square (i.e., Eq. (1)) robust, the authors used the least median squares which minimise Eq. (2):

$$\mathrm{Median}_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2 \tag{2}$$

Hawkins et al. [31] model the outlier detection problem as a hybrid of regression analysis and neural networks. The number of inputs and outputs of the neural network is made equal to the number of attributes of the problem being examined. The algorithm learns from the data using the neural network. The total error imposed by recovering each data object from the neural network is computed using the regression model. Data objects with the maximum recovery error are considered outliers. Roberts uses extreme value theory (EVT) with the Gaussian mixture model to represent data of abnormally low or high value in the tails for outlier detection [60]. Extreme value theory examines the tails of the data distribution and then estimates the probability that a given instance of data is an extreme value in the distribution. The probability of observing some extreme value in the model is given by Eq. (3):

$$P(\textit{extreme}_x) = \exp\left\{-\exp\left(-\frac{x_m - \mu_m}{\sigma_m}\right)\right\} \tag{3}$$

where μ_m, σ_m are normally defined as the norming parameters that depend only on the sample size m and are estimated using heuristics. Miller et al. [52] stated, given any two constants r and c, μ_m, and σ_m can be computed as follows:

$$\sigma_m = \sigma = c \tag{4}$$

$$\mu_m = \sigma \ln(m) = r \tag{5}$$

Extreme value theory is ideal for outlier detection because it is not limited by previously seen classes as happens with classification algorithms. It is ideal for situations where negative examples are very hard to obtain in real life situations [26,60].

Bi et al. [16] propose the Discrete Gaussian Exponential 'DGX' as a new probability distribution for fitting data that are massively skewed. The experimental results with many real-world data including sales data from a large retailer chain, age data from AT & T, and web usage data reveal the distribution fits those data very well with a correlation of about 99% when investigated. It is worth noting that outliers are a subset of massive skewed data in real life. The mean and variance of 'DGX' can provide valuable information for clustering and outlier detection unlike other discrete distribution where the distribution parameters (i.e., mean, median, variance, etc) are much less valuable [16].

2.1.1. Merits of distribution-based techniques

The distribution-based techniques are based on models with sound mathematical backgrounds. For example, hypothesis verification is a conventional problem in mathematical statistics so its application to outlier detection is mathematically justifiable. Secondly, the distribution-based techniques are very efficient and provide outliers that can be meaningfully explained if the data distribution is known. Thirdly, they are easy to use since the miner does not have to understand the rigorous mathematics behind a particular distribution. The majority of the distributions are summarised in tables that are very easy to use. Finally, the techniques grow with the complexity of the models and not with the data size which makes them more scalable.

2.1.2. Shortcomings of distribution-based techniques

The availability of too many discordancy tests makes the outlier detection problem very elusive and time consuming. The few tests available for multidimensional data are very complex to understand and often very difficult to use [42]. Thirdly, the miner is required to know the data distribution before a test can be applied. This is particularly difficult because in a typical data-mining environment it is almost impossible to assign data to a specific distribution. Data are usually multidimensional and may not follow any single distribution (e.g., normal, gamma, etc). Finally, the construction of hypothesis is nontrivial for complex distributions [28].

2.2. Depth-based techniques

The depth-based methods developed from computational statistics do not fit data points with statistical distributions. They represent data objects in space, and assign a depth to each object based on some definition of depth. Depth in a two dimensional plane is defined as follows:

Definition 1. Let $\psi = \{\psi_1, \psi_2, \ldots, \psi_n\}$ be a finite set of points in R^2 (i.e., 2-dimensional plane) and λ be an arbitrary point in ψ. The depth of λ with respect to ψ in R^2 is the minimum number of points of ψ in any closed half-plane bounded by a line that passes through λ. □

Data objects that appear in shallow layers are expected to contain more outlying data than those that appear in deep layers [59,73]. A robust notion of depth called depth contours operates under the assumption that a point p in a space is of depth k if k is the minimum number of points that must be removed to expose p. The k-th depth contour marks a boundary between all points with depth k and all those with depth $k + 1$. ISODEPTH [61] and Fast Depth Contours (FDC) [37] are two depth-based algorithms based on depth contours. ISODEPTH relies on the computation of dividers and also spend

time removing collinear point, which renders it very slow. FDC on the other hand is resistant to collinear points and restricts divider computation to only selected subsets of the data. Hence, FDC is more efficient than ISODEPTH.

The *grand tour* algorithm rotates the space containing the data points geometrically several hundred times while noting the position of every objects each time. Non-outlying objects will appear within an ellipse while the outlier candidates' always fall outside the ellipse [12]. Bartkowiak et al. [21] examine the outlier candidates generated from *grand tour* using regression analysis. Closely related to the *grand tour* is the Minimum Volume Ellipsoid estimation (MVE), which fits data points with the smallest possible ellipsoid around the majority of the data. Data points falling outside the volume are declared outliers [59]. Similarly, convex peeling removes data falling on the boundaries of the convex hull of the data distribution. The major difference is that MVE defines a volume around all the data points while convex peeling assigns depth to individual data and removes those with shallow depths as outliers.

Deviation detection using implicit redundancy has been somehow arbitrarily placed under depth-based techniques for convenience. Arning and Agrawal [1] propose a technique for determining deviation based on the dissimilarity within the dataset called *sequential exception*. Given any set of interest, $\{\mathrm{I}\}$, and all possible subsets of $\{\mathrm{I}\}$ called the power set $\mathrm{P}(\{\mathrm{I}\})$, the item(s), which if removed from the power set, contributes most to the dissimilarity of the original set with the least number of items constitutes *an exception set.* The algorithm requires a function that gives the degree to which elements in a data set causes the dissimilarity of the data set to increase is called *implicit redundancy*. The *implicit redundancy* techniques declare a set of objects to be outliers if their removal causes the remaining objects within the set to be more homogenous [1].

2.2.1. Merits of depth-based techniques

The depth-based techniques do not require a priori knowledge of any statistical distribution. Thus, the depth-based techniques are supposed to address the problem of distribution fitting and conceptually support outlier detection from high dimensional data.

2.2.2. Shortcomings of depth-based techniques

Theoretically, the depth-based techniques should be computationally unrestricted in the number of dimensions. However, in practice, they become inefficient for high dimensional data (i.e., $k \geqslant 3$) because they rely on the computation of a $k-d$ convex hull, which has lower bound complexity of $\Omega(\mathrm{N}^{\lfloor k/2 \rfloor})$ for n objects, where k is the number of dimensions and N is the data size. Secondly, the concept of depth or volume is very difficult to define for high dimensional data. Hence, the depth-based techniques are not capable of addressing the dimensionality curse of the distribution-based techniques.

2.3. Distance-based techniques

The distance-based outlier concept started as the unified notion of outliers [42]. The concept describes an object O in a dataset T as a *DB*(p, D)-outlier if at least fraction p of the objects in T lies greater than distance D from O [43]. This definition captures the general idea of outliers described by Hawkins [27], and unifies most of the distribution-based notions of outliers. The choice of the parameters p and D are left to the miner to define since p and D may vary for different applications. Knorr et al. [42,43,45] show that if an object S is an outlier according to a specific discordancy test, then S is also a *DB*(p, D)-*outlier* for some specified p and D.

The nested loop, index-based, and the cell-based algorithms are proposed for mining all *DB*(p, D)-outliers. The nested loop and index-based algorithms have quadratic time complexity because they

involved pair wise distance computations for every data object before the top-n candidates with the largest distances are selected as outliers. The index-based algorithm is less popular because of the high cost associated with building indexes for participating dataset [43]. The cell-based algorithm, with a linear time complexity, partitions the space into cells and constructs estimates for pair-wise distances between objects which reduce the complexity of examining all pairs.

Ramaswamy et al. [62] propose a new definition for outliers based on the distance of a point to its k^{th} nearest neighbor that allows outliers to be ranked. Outliers are described as top-n data whose distance to the k^{th} nearest neighbor is the greatest. 'Given a k and n, a point p is an outlier if no more than n-1 other points in the dataset have a higher value of D^k than n, where D^k denote the distance of the k^{th} nearest neighbor to a point' [62]. The top-n points with the maximum D^k value are declared outliers. The miner is only required to specify the number of outliers needed without specifying the parameters p and the neighborhood distance D. The partition-based algorithm that uses linear time clustering to partition the space into regions to facilitate faster identification of neighbors is proposed. Summary statistics such as minimum bounding rectangle (MBR) are stored for each region and used during the nearest neighbor search. Data points are compared to the MBR to determine if it is possible for a nearest neighbor to come from that region. If not all the points in the region are eliminated. The partition-based algorithm performs better than the nested loop and the index-based algorithms on low dimensional synthetic data [62].

Hung and Cheung propose an efficient version of the nested loop (*NL*) algorithm called enhanced nested loop (*ENL*) and its parallel version (*PENL*) [29]. The *ENL* and *PENL* algorithms are linear to the number of dimensions and data size. The experimental results using the Bulk Synchronization Parallel (BSP) model reveal *PENL* outperforms *ENL*. The improvement in performance is because *PENL* is linear to the number of processor used. A further test using IBM's parallel computer system (i.e., IBM 9076 SP2) shows *PENL* is a better choice for mining outliers from a cluster of workstations.

Shekhar et al. [66] propose methods for mining graph-based outliers. Their approach examines the neighborhood of points topologically connected together. They describe a node in the graph to be an outlier if the difference between its sensor value and the average sensor value of its topological neighbors differs by more than some threshold from the mean difference between all nodes and their topological neighbors. The algorithms applied to the twin cities traffic data gave very interesting outliers. The graph-based algorithm is applicable to domains where connected graph of nodes can be established such as analyzing network traffic flow.

Anguilli and Pizzuti [9,10] define a new distance-based outlier that takes into account weighted neighborhood distances of each object. Outliers are defined as data objects with larger weights computed using the *Hilbert space filling curve to* project data over the interval [0, 1] multiple times. The estimate of an outlier score is improved over the full feature space at each successive projection. The algorithm scales almost linearly in both dimensionality and data size but the results reported are for only two synthetic datasets. Their approach has yet been tested on real-world datasets. Secondly, the performance of the algorithm is not compared with any of the existing distance-based outlier mining algorithms.

The basic nested loop algorithm has a quadratic time complexity [43,62]. To address this problem, Bay and Schwabacher [22] modify the original nested loop algorithm to yield near linear time complexity in conjunction with randomization and simple basic pruning. For each data D, its closet neighbors found so far are kept; D is removed if its closest neighbors achieve a score less than the outlier *cut-off* because it can no longer be an outlier candidate. The algorithm finds more extreme outliers as more data are processed and the outlier *cut-off* increases with the pruning efficiency. The results show the algorithm has the best scaling for distance-based algorithms with a near linear time complexity on large real-world dataset with millions of records and many features. However, it is not known how this algorithm compares to *ENL*

and *PENL* [29]. Recall that *ENL* and *PENL* are the enhanced and parallel versions of the original nested loop algorithm, respectively.

Breunig et al. [18,19] introduce the *local outlier* concept and argue being an outlier is not just a binary property as entailed in the other distance-based concepts. They contend that for many scenarios it is more appropriate to assign a degree of outlying called the *local outlier factor* that depends on the remoteness of an object to its surrounding neighborhood. Data objects with high *local outlier factors* have higher tendency of being outliers than those with low *local outlier factors*. It is further agued that taking a global view of the entire datasets, as done in the earlier algorithms (e.g. [42,43,62]), produces global outliers [19]. They state that for datasets with more complex structures, some data may be outliers within their neighborhood but not outliers in the entire dataset. The *LOF* algorithm that computes the local outlier factor values (LOFvalues) for every object in the dataset is proposed. The results from the National Hockey League player statistics for 1996, the German National Soccer League 1998/99, and other synthetic data demonstrate the superiority of *LOF* algorithm over the nested-loop and index-based algorithms. The time complexity is similar to that of the nested-loop and index-based algorithms (i.e., quadratic).

Jin et al. [38] propose a micro-cluster-based algorithm for mining *top-n* local outliers. The algorithm restricts *LOFvalue* computation to only selected clusters that constitute the candidate outlier set. It first computes the lower and upper limits for each cluster based on their local reachability densities. The limits are used to determine which of the clusters could possibly contain outliers. The clusters that cannot possibly contain outliers are eliminated from the possible outlier candidate set. Pruning of the clusters reduces the number of *LOFvalue* computations to only data in the candidate set. Though more efficient than the original local outlier factor algorithm, it still relies on computing reachability distances and local reachability densities for every object in the dataset before pruning.

Tang et al. [72] combine the k-nearest neighbor approach with a connectivity-based graph that computes the weighted distance score of each object. Their algorithm computes the average chaining distance of a point *p to* its $k-$nearest neighbors by assigning higher weights to points in sparse regions. Points in sparse regions have relatively high average chaining distances and hence are more likely to be outliers. The algorithm incorporates density and isolation in that a point may lie in a relatively sparse region without necessarily being an outlier. However, an isolated point is an automatic outlier. The downside is that the approach has almost the same runtime complexity as the k-nearest neighbor algorithms.

Agyemang et al. [5] propose the *LSC-Mine* algorithm based on the original idea of local outliers but *LSC-Mine* avoids computing reachability distances and local reachability densities for every object in the dataset. Instead local sparsity ratios derived from the neighborhood distances are computed. The algorithm prunes data objects that are not possible outlier candidates using a pruning factor computed from the neighborhood distances. The next phase of the *LSC-Mine algorithm* computes local sparsity coefficients for objects in the candidate set. Data objects with high local sparsity coefficients have a higher tendency to be outliers than those with low local sparsity coefficients. Since outliers constitute only a small fraction of the entire dataset, the pruning in *LSC-Mine* is able to remove about half of the dataset as non-outlier candidates.

Recently, Aggarwal and Yu [13,14] argue most of the existing outlier mining techniques are based on methods that make implicit assumption of relatively low dimensionality of the data containing the outliers. They argue such implied assumption renders the algorithms ineffective for high dimensional datasets with sparse distribution and propose new technique for mining outliers in high dimensional space [14]. Their techniques detect outliers by observing the density distributions of data projections. Their algorithms declare data points that appear in a local region of abnormally low density in some

lower dimensional projections as outliers [13]. The basic idea involves defining outliers by examining those projections of data with abnormally low density. Two algorithms, a naïve and an evolutionary algorithm, are proposed for mining such high dimensional outliers.

The naïve algorithm examines all subsets of dimensions for patterns with abnormally low presence which makes it extremely slow. The sparse patterns are later used to determine data points that are possible outliers. The evolutionary algorithm on the other hand is capable of finding hidden combination of dimensions with sparse patterns without searching the entire space [14]. The results with synthetic and real datasets reveals finding outliers using lower dimensional projections overcome the curse of dimensionality embodied in the distance-based algorithms. In other words, the evolutionary algorithms scale up to hundreds of dimensions and data size in the order of 10^5 compared to less than 10 dimensions for the distance-based algorithms [13,14].

2.3.1. Merits of distance-based techniques

The distance-based techniques are independent of any probabilistic models and hence do not require *a priori* knowledge of the data distribution. In addition, the distance-based algorithms scale very well for huge multidimensional datasets; at least better than the depth-based techniques. The algorithmic parameters are very easy to choose and the results are easy to interpret. The local outlier concept is able to capture different outlier varieties that are missed by the earlier algorithms.

2.3.2. Shortcomings of distance-based techniques

Almost all the distance-based techniques have quadratic time complexity. The efficiency can be improved by applying early pruning techniques to reduce the outlier candidate set [22]. Secondly, they are dependent on parameters like the minimum outlier distance D, and the number of neighbors k, which tend to affect their performance if chosen wrongly. Finally, the distance-based techniques are not applicable to situations where the notion of distance cannot be easily established. Hence, they may not be applicable for mining outliers from text datasets and web repositories directly. The distance-based techniques have to be modified if they are to be used for mining outliers from text and web data repositories [2].

3. Cluster-based techniques

The cluster-based techniques are identified as a partial class on the taxonomy because they are not fully outlier mining techniques. They are clustering algorithms designed in the context of data mining with exception/outlier handling capabilities (e.g., CLIQUE [6], DenClue [32], WaveCluster [63], STING [76], BIRCH [80]). The outliers are generated as a by-product of the clustering. Recently, support vector clustering algorithms have been suggested for mining outliers. A support vector machine is a hyperplane that separates a set of positive data from a set of negative data with maximum margin in the feature space [68,78]. Data objects that lie on the boundaries of the margin called support vectors are considered normal while those that lie outside the margins are considered outliers [17,35,54]. These algorithms will not be discussed in detail because their main objective is to find clusters and hence they are optimized for clustering and not for outlier mining.

It is worth noting that some distance-based and density-based algorithms (e.g. [38,62], etc.) use clustering algorithms to achieve their goals. The difference however is that where the latter is aimed at finding outliers the former attempts to find clusters. Similarly, clustering algorithms may either use numeric or non-numeric data. The classification in the taxonomy is more meaningful in the context of outlier mining.

3.1. Merits of cluster-based techniques

The only advantage of the cluster-based techniques is that they generate clusters in addition to the outliers they produce as a by-product. Recall that these algorithms are optimized for finding clusters and not outliers.

3.2. Shortcomings of cluster-based techniques

The notion of outliers in the cluster-based techniques is essentially binary, which means an object is either an outlier or a not. They do not provide any means of ranking the outliers produced. Secondly, the clustering algorithms consider any data falling outside a cluster as an outlier, which may not necessarily be true in a data mining context. Finally, the clustering algorithms do not distinguish between noise and outliers.

4. Symbolic techniques

The symbolic techniques consist of outlier mining techniques developed outside clustering and numeric domains for finding outliers. They consist of web-based techniques and rule-based techniques. The current web-based techniques (e.g. [2,3]) use the contents of web pages and the HTML tags to defined dissimilarity among documents. The text-based techniques may be considered as a special case of the web-based methods where only the texts on the documents are used (e.g. [30,46]). The rule-based techniques extract useful rules and present a subset of such rules to the user as exceptions or surprising patterns based on the statistical properties of the discovered rules.

4.1. Web-based techniques

The web-based techniques are designed for finding outliers from web data repositories containing data of multiple types (e.g., text, hypertext, hyperlinks, video, etc.). Liu et al. [50] propose algorithms for finding unexpected information from a competitor's website. Their approach focuses on finding competitors using only information contained in the web contents. The proposed algorithms are very successful when applied to real-world datasets. However, the approach requires a query vector before the mining process begins. It also requires a competitor's website to be identified *a prior* meaning data with unidentified domains cannot be mined. Finally, the Liu et al. [50] approach is dedicated to finding interesting and unexpected patterns from a competitor's website, which may not necessarily be outlying patterns. In other application, Jung et al. [36] used semantic outlier mining techniques for sessionalizing web log data to track the behaviour of individuals who may easily change their interests and intensions while surfing the web.

Agyemang et al. [2,3] propose a definition for the web outlier concept. They define various types of outliers available on the web and a taxonomy of web outliers. The authors propose the WCOW-Mine algorithm for mining web content outliers using keywords, HTML structure, and a domain dictionary. The WCON-Mine algorithm is as an n-gram alternative for mining web content outliers [3]. These two algorithms compare each document with a domain dictionary and those that tend to be very dissimilar to the dictionary are declared outliers. The good news is that both algorithms are able to identify outliers from the web contents of web documents. The downside is that both algorithms rely on domain dictionaries that may require the services of a domain expert which can be error-prone and costly. The

HyCOQ algorithm mines web content outliers using hybrid data without a domain dictionary [4]. The algorithm uses IR techniques to extract useful features from web documents and then applies dissimilarity algorithms to determine outlying documents based on the computed nearest dissimilarity density defined below:

Definition 2. Let d be a document, $Neigh_k(d)$ the neighborhood dissimilarity of d, and DIS be the document dissimilarities. The nearest dissimilarity density of d, denoted $Nearest_k(d)$, is defined as the ratio of the cardinality of neighborhood dissimilarity of d to the sum of the actual document dissimilarities within the neighborhood. □

Documents with high nearest dissimilarity densities are more likely to be outlying than those with low nearest dissimilarity densities. The experimental results using embedded motifs show HyCOQ is more accurate than WCOW-Mine and WCON-Mine on the *webkb* [75] dataset. However, the overall response time for WCOW-Mine and WCON-Mine algorithms are better than HyCOQ algorithm [4].

In a special case when only the text contents of the web documents are of interest, the web-based techniques may be applied without the weighting functions for the different HTML tags (refer to [2,3] for details). Keila and Skillicorn [46] apply a deceptive model using Single Value Decomposition (SVD) to the Enron email dataset, a public dataset of emails to and from Enron staff, in a three year period before the collapse of Enron. They ranked the email based on how well they fit the deceptive profile. Their method identified emails from the top executives as suspicious. It also raised flag about other potential problems in the organization such as complaining and spending of organizational resources on non-work related issues [46]. These flags are good signatures that can be analysed further for fraud, criminal behaviour, and other organizational malpractices.

Security authorities use keyword filtering to select messages from a set of intercepted messages to be analysed further for fraud. Replacing words on keyword list with different words with the view of eluding authorities is called a keyword replacement strategy. Skillicorn shows that fraudulent conversations are detectable even if fraudsters replace suspected keywords that may attract the attention of authorities with other words (i.e., words that may not be found on the suspected keyword list) in their conversation [65]. He argues that replacement strategy creates a signature in the altered messages that make them easily detectable using several forms of matrix decomposition. He used Singular Value Decomposition (SVD) to analyse a real-world dataset and concludes that not only can such fraudulent conversations be detected but also a set of related messages are detectable as well even if the source is not static [65].

Miller and Myers [51] viewed outlier mining as a new way to reduce errors by focusing users' attention on possible errors in text processing. The search-and-replace command in text processors forces users to replace a match at a time with confirmation or replace all matches at once. They argue that none of the two choices is ideal for long documents. Confirming a match is tedious and error-prone because if most of the match is YES, users tend to be pressing YES without even checking if there is actually a match. Similarly, replacing all matches without confirmation means the user should trust the precision of the match patterns, which cannot be guaranteed.

Miller et al. [51] propose an outlier finder with the search-and-replace command in a text editor. The text processor highlights only the most outlying pattern matches, which enable the user to focus only on matches that are problematic. The algorithm focuses on finding errors in text data using substrings of texts from documents as inputs. It compares each substring with other substrings from the documents. Similarities are computed from weights assigned to substrings. Substrings that are more correlated receive higher weights while uncorrelated substring receives lower weights and are eventually declared mismatches and hence outliers. The experimental results show the approach enhances performance and

reduces errors due to the search-and-replace command. The major drawback is that this algorithm is too specialised and cannot be used for any major outlier mining tasks apart from the search-and-replace task in text editors.

He et al. [30] claim that without using the *semantic knowledge* of the underlying data, the outliers produced cannot be meaningfully interpreted and hence cannot be meaningful outliers. The issue of semantics is resolved by proposing a new definition for outliers that accounts for the semantic meaning of data called a *semantic outlier*. A semantic outlier is described as a data point that behaves differently with other data points in the same class [30]. The degree of semantic outlying called the semantic outlier factor is established. The algorithm for finding all semantic outliers called, *FindSOF*, initially partitions the data into clusters. The *FindSOF* algorithm computes pair-wise record similarities within clusters during the first scan and then among clusters during the second scan and returns the outliers based on their similarities.

The results from 1984 United States Congressional voting reveal some interesting patterns at the senate. The results on 16 issues at the senate floor show a Republican who votes on 8 issues like a Democrat and 5 issues like a Republican. Thus, the senator is more of a Democrat than a Republican as far as the 16 issues on the senate floor is concerned. They conclude that *FindSOF* is capable of finding interesting outliers but remark that it may compliment existing outlier mining algorithms [30]. The drawbacks of *FindSOF are* (1) it only works with categorical data, and (2) the dataset used for the experiment is too small to draw any meaningful conclusions. Additionally, the results from *FindSOF* are not compared with any algorithm. The dataset had 435 records made up 168 Republicans and 267 Democrats and their votes on 16 issues.

4.1.1. Merits of web-based techniques

The web-based techniques have been capable of finding outliers from web documents. They have also been shown to be very effective when documents containing only text are considered. Kaila et al. [46] used SVD to identify deceptive communication in emails. The web-based techniques allow documents to be ranked based on either their dissimilarity to a dictionary or other documents in the domain.

4.1.2. Shortcomings of web-based techniques

The web-based techniques consider documents with text contents only or text and HTML tags containing text. These web-based algorithms do not exploit other data types such as video, audio, hyperlink, and other rich data types that may hold valuable information that distinguish one document from the other. Most of the web-based algorithms have quadratic time complexity like their numeric-based counterparts.

4.2. Rule-based techniques

Rule-based techniques can generally be grouped into rule discovery and rule learning. Rule discovery usually examines many candidate rules whereas rule leaning deals with smaller set of rules [25]. This survey treats rule discovery and rule learning as synonyms. A rule is a statement of the form ‘if premise then conclusion’ [8]. Rules are described as strong or weak based on confidence and support. A rule may be described as strong, in which case it describes regularity for many objects or weak, if it applies to fewer objects. Interesting and surprising rules are generally fewer and are often found among the weak rules [67]. This survey reviews some of the techniques design for finding interesting and unexpected or exception rules. Techniques for exception-rule discovery are classified into two namely: (1) directed approach when the technique is provided with background knowledge usually in the form of rules; (2) undirected approach when no background information is provided [67].

Liu et al. [48] propose a technique called the user-expectation method in which the user is first asked to provide some expected patterns according to past knowledge or intuition about the data. Their technique matches the user-expectations with the discovered patterns using fuzzy matching algorithms and results are ranked accordingly. Their algorithm is very general, interactive, and can perform different ranking depending on the user's input. The rules are ranked according to their unexpectedness (i.e., patterns are interesting if they can 'surprise' the user) [70]. Their algorithm does not study actionability (i.e., a rule is interesting if the user can do something with it) [55].

Measures of interestingness of patterns in data mining applications can be classified into objective (i.e., those that depend only on the structure of a pattern and the underlying data used in the discovery process) and subjective (i.e., those that depend on the class of users who examine the pattern). Silberschatz and Tuzhilin [70,71] classified subjective measures into *actionable* and *unexpected* patterns and argue that the two are independent of each other (i.e., actionable patterns are unexpected and most unexpected patterns are actionable). They concentrated on unexpectedness as a tool for measuring interestingness. Silberschatz et al. [71] defined interestingness as strong rules that 'changes' the belief system. They provide techniques based on the Bayes theorem, statistics, and the frequency of the rules, for measuring interestingness without any experimental analysis.

Padmanabhan et al. [56–58] propose characterization of *unexpectedness* based on a system of prior beliefs. Their algorithms [56,57] focus on finding unexpected patterns while others [58] deal with the problem of finding minimal patterns. The latter work combines two independent concepts; *minimality* and *unexpectedness* into an integrated concept that identifies small but interesting patterns. The experimental results with real-world data shows the MinZoomUR algorithm (i.e., proposed algorithm) discovers fewer patterns than comparable algorithms and yet retains most of the truly interesting and unexpected patterns. Jaroszewicz and Scheffer propose similar methods using sampling-based algorithms and the Bayesian network for finding unexpected patterns from large datasets with help of background knowledge [40].

Hussain et al. [33] propose an undirected method that exploits the knowledge extracted from data for measuring interestingness. The measure is unbiased because the knowledge comes from the data and it is not polluted with any user beliefs. In their work they used the extracted rules, confidence, and support to define relative interestingness (RI). They argue that if no other information is given then a rule with low confidence gives more information that a rule with higher confidence. Two sets of interestingness are distinguished; confidence-based interestingness and support-based interesting. Thus, given any rule $AB => X$ relative to $A =>$X, $B => X$, the relative interestingness of the rule (*RI*) is the combination of relative interestingness due to support (RI_S) and relative interestingness due to confidence (RI_C) (i.e., $RI = RI_S^{AB} + RI_C^{AB}$) [33].

Liu et al. [49] propose a method for summing the rules discovered from association rule mining using a subset of the unpruned rules called Direct Search (DS). The DS uses statistical correlation as the basis for finding rules that represent the fundamental relations of the domain instead of minimum confidence. This allows more meaningful and interesting rules to be identified easily. The significance of a rule is measured using the chi-square test for correlation from statistics. The experimental results with real-world dataset show the set of rules generated from DS are very small which can be analysed by humans to identify the most interesting and useful rules.

Suzuki provides a very good summery of research on undirected methods for exception rule mining [67]. The paper provides a good reference points for readers interested in interestingness and unexpectedness of rules. It concentrates on discovering a set of interesting rule pairs from data sets. The algorithms assume the existence of an evaluation criteria or a set of possible constraints. It provides

MEPRO for discovering rule pairs based on the interestingness measure J [64]. The J-measure for a rule $x => y$ is given by:

$$J(x;y) = \Pr(y)\ j(x;y) \tag{6}$$

$$j(x;y) = \Pr(x/y)\log_2\ \frac{\Pr(x/y)}{\Pr(x)} + \Pr(\bar{x}/y)\log_2\ \frac{\Pr(\bar{x}/y)}{\Pr(\bar{x})} \tag{7}$$

The interestingness of the rule pair is defined as the product of $ACEP(x, Y_\mu, x', Z_v)$ of J-measure of strong rule and J-measure of an exception rule and is given by:

$$ACEP(x, Y_\mu, x', Z_v) \equiv J(x; Y_\mu)\ J(x'; Y_\mu \wedge Z_\nu) \tag{8}\ [67, 69]$$

PADRE is proposed to evaluate the reliability of the rule pairs generated. Techniques for estimating the thresholds used in PADRE are provided. The experimental evaluation of MEPRO and PADRE on real-world data reveal both algorithms are very efficient and effective in determining a set of interesting and unexpected rule pairs from data [67].

4.2.1. Merits of rule-based techniques

The directed methods are likely to produce smaller and stronger rules because most irrelevant rules can be removed with the help of the domain knowledge. The undirected methods can produce more relevant and unbiased results based on only the rules without help from any expert but this can be very costly.

4.2.2. Shortcomings of rule-based techniques

The directed methods rely on domain knowledge in the form of exception rules, which require the services of a domain expert. Creating the domain knowledge may be error-prone and biased due to incomplete or inaccurate knowledge about the domain. On the other hand, the undirected methods are also expensive in terms of computational cost. Compared to the directed methods, the undirected methods suffer from extra search overhead for strong rules [33,67].

5. Future research directions

The nonparametric methods do not rely on distribution parameters and hence are a very good substitute for the distribution-based techniques. However, the non parametric methods are not useful for large sample sizes. Researching and developing non-parametric methods that scale for large sample sizes is an open research issue. Another area will be developing new distributions for skewed data with emphasis on outlier detection.

The popularity of the World Wide Web has attracted a great number of companies to do business on-line. In view of this, more efficient techniques for detecting fraud and other criminal activities on the web are required. Unfortunately, the electronic information is not only numeric which makes it extremely difficult to detect web-based outliers (fraudulent and other criminal activities) directly with existing numeric outlier mining techniques. Improving existing numeric outlier mining techniques to accept non-numeric data is a viable research issue. Another critical research issue is improving the scalability of outlier mining algorithms because most are quadratic. The time complexity may be

improved by introducing more refined and rougher models into the existing outlier detection algorithms (e.g., introduction of micro-clusters and early pruning [38] and the use of kennel functions [54]).

There are several benchmark data for testing the performance of numeric outlier mining algorithms but there is none for testing the performance of web-based algorithms. Establishing benchmark data for testing web-based algorithms may help researchers to evaluate the performance of their algorithms on common data. Moreover, it is important to evaluate the performance of numeric algorithms applied directly to mine symbolic data and to compare the performance to results from symbolic outlier mining algorithm. Existing web-based algorithms use either text alone or text with HTML tags, ignoring the other rich data types available in that domain (e.g., image, audio, hyperlinks, etc). Research exploring more of these data types is very interesting. Finally, a growing number of web pages are based on *XML-schema* and hence it is very important to integrate them into the algorithmic design.

The application of outlier mining techniques to life sciences is currently a key research issue. The challenge is that the nature of the data means most algorithms discussed here cannot be applied directly without modification because an outlier must be defined on a case by case basis. The definition of an outlier may involve defining a criteria based on which each data can be determined whether it belongs to the majority of the cases or not. If a specific data item belongs to the majority, then it is normal and an outlier otherwise (with respect to specific criteria). How the criteria is defined affect the type of outliers. The challenge lies in defining the outlier models suited for a particular life science domain. Generally, defining outlier models is not trivial but identifying outliers based on the models is very simple and trivial

6. Conclusion

Data that do not obey rules considered normal for the majority of the data elements are referred to as outliers. In other words, outliers are considered to be the odd ones in the mist of data. Despite their odd characteristics outlier mining has very important application in areas such as intrusion detection, fraud detection, and the identification of competitors in electronic commerce among many others. There are several methods designed for mining outliers from different domains. In this survey, we provide a taxonomy for outlier mining techniques and place the techniques under their appropriate headings.

This survey identifies numeric, non-numeric and cluster-based methods as three major outlier mining techniques. The numeric techniques consist of all outlier mining techniques that use numeric data directly aside from cluster-based techniques. The cluster-based techniques are clustering algorithms with outlier handling capabilities. They are identified as partial outlier mining class because such techniques are optimized for clustering rather than outliers. The cluster-based techniques may either use numeric or non-numeric data as input. The symbolic techniques, on the other hand, uses non- numeric data and include web-based and rule-based techniques.

In addition, the survey provides some practical applications for outlier mining techniques and discusses the strengths and the weaknesses of the techniques provided. It also provides insight into some of the research issues in the area. In conclusion, this survey provides a comprehensive review of outlier mining techniques from several domains. It provides a good starting point for researchers interested in outlier mining.

References

[1] A. Arning, R. Agrawal and P. Raghavan, *A Linear Method for Deviation Detection in Large Databases*, Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, 1996, 164–169.

[2] M. Agyemang, K. Barker and R. Alhajj, *Framework for Mining Web Content Outliers*' Proceedings of the 19th ACM Symposium on Applied Computing, Nicosia, Cyprus, 2004, 590–594.
[3] M. Agyemang, K. Barker and R. Alhajj, *Mining Web Content Outliers Using Structure Oriented Weighting Techniques and N-grams*, Proceedings of the 20th ACM International Symposium on Applied Computing, Santa Fe, New Mexico, USA, 2005, 482–487.
[4] M. Agyemang, K. Barker and R. Alhajj, *Hybrid Approach to Web Content Outlier Mining Without Query Vector*, Proceedings of the 7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK), LNCS-3589, Denmark, 2005, 285–294.
[5] M. Agyemang and C.I. Ezeife, *LSC-Mine: Algorithm for Mining Local Outliers*, (Vol. 1), Proceedings of the 15th Information Resource Management Association (IRMA) International Conference, New Orleans, USA, 2004, 5–8.
[6] R. Agrawal, J. Gehrke, D. Gunopulos and P. Raghavan, *Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications*, Proceedings of ACM SIGMOD International Conference on Management of Data, Seattle, WA, 1998, 94–105.
[7] R. Agrawal, T. Imielinski and A. Swami, Data Mining: A Performance Perspective, *IEEE Transactions on Knowledge and Data Engineering* **5**(6) (1993), 914–925.
[8] R. Agrawal, T. Imielinski and A. Swami, Mining association rules between sets of items in large databases, *ACM SIGMOD Records* **22**(2) (1993), 207–216.
[9] F. Anguilli and C. Pizzuti, in: *Fast Outlier Detection in High Dimensional Spaces*, T. Elomaa, ed., PKDD, LNAI 2431, 2002, pp. 15–27.
[10] F. Anguilli and C. Pizzuti, Outlier Mining in Large High-Dimensional Data Sets, *IEEE Transactions on Knowledge and Data Engineering* **12**(2) (2005), 203–215.
[11] A. Adam, E. Rivlin and I. Shimshoni, ROR: Rejection of Outliers by Rotation, *IEEE Transaction on Pattern Analysis and Machine Intelligence* **23**(1) (2001), 78–84.
[12] D. Asimov, The grand your: a yool for viewing multidimensional data, *SIAM J.Sci. Stat. Compu* **6** (1985), 128–143.
[13] C.C. Aggarwal and P.S. Yu, *Outlier Detection for High Dimensional Data*, Proceedings of ACM SIGMOD International Conference on Management of Data, Santa Barbara, CA, USA, 2001, 37–46.
[14] C.C. Aggarwal and P.S. Yu, An effective and efficient algorithm for high-dimensional outlier detection, *The VLDB Journal* **14**(2) (2005), 211–221.
[15] R.J. Bolton D.J. Hand, *Unsupervised Profiling Methods for Fraud Detection*, In Conference of Credit Scoring and Credit Control VII, UK September 5–7, 2001.
[16] Z. Bi, C. Faloutsos and F. Korn, *The "DGX" Distribution for Mining Massive Skewed Data*, Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2001, 17–26.
[17] A. Ben-Hur, D. Horn, H.T. Siegelmann and V. Vapnik, Support vector clustering, *Journal of Machine Learning Research* **2** (2001), 125–137.
[18] M.M. Breunig, H.-P. Kriegel, R.T. Ng and J. Sander, *OPTICS-OF: Identifying Local Outliers*, Proceedings of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Czech Republic, (LNAI 1704), 1999, 262–270.
[19] M.M. Breunig, H.-P. Kriegel, R.T. Ng and J. Sander, LOF: identifying outliers in large dataset, *Proceedings of the ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA* **29**(2) (2000), 93–104.
[20] V. Barnett and T. Lewis, *Outliers in Statistical Data*, John Willey, 1994.
[21] A. Bartkowiak and A. Szustalewicz, Detecting utliers by a grand tour, *Machine Graphics and Vision* **6** (1997), 487–505.
[22] D.S Bay and M. Schwabacher, *Mining Distance-Based Outliers in Near Linear Time with Randomization and Simple Pruning Rule*, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery, Washington, DC, USA , 2003, 29–38.
[23] P. Chan, W. Fan, A.L. Prodromidis and S.J. Stolfo, *Distributed Data Mining in Credit Card Fraud Detection*, IEEE Intelligent Systems, Nov.-Dec., 1999, 67–74.
[24] U. Fayyad, G. Piatesky-Shapiro and P. Smyth, *Knowledge Discovery and Data Mining: Towards a Unifying Framework*, Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, 1996, 82–88.
[25] J. Furnkranz, Separate-and-Conquer Rule Learning, *Artificial Intelligence Review* **13** (1999), 3–54.
[26] V.A. Hodge and J. Austin, A survey of outlier detection methodologies, *Artificial Intelligence Review* **28** (2004), 85–126.
[27] D. Hawkins, *Identification of Outliers*, Chapman and Hall, London, 1980.
[28] A.S. Hadi, A new measure of overall potential influence in linear regression, *Computation Statistics Data Analysis* **14** (1992), 1–27.
[29] E. Hung and D.W. Cheung, Parallel algorithms for mining outliers in large databases, *Distributed and Parallel Databases* **12**(1) (2002), 5–26.
[30] Z. He, S. Deng and X. Xu, in: *Outlier Detection Integrating Semantic Knowledge*, X. Meng, J. Su and Y. Wang, eds, WAIM 2002, LNCS 2419, 2002, pp. 126–131.

[31] S. Hawkins, H. He, G. Williams and R. Baxter, *Outlier Detection Using Replicator Neural Networks*, Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery, 2002, 170–180.
[32] A. Hinneburg and D.A. Keim, *An Efficient Approach to Clustering in Large Multimedia Databases with Noise*, Proceedings of 4th International Conference on Knowledge Discovery and Data Mining, New York City, NY, 1998, 58–65.
[33] F. Hussain, H. Liu, E. Suzuki and H. Lu, *Exception Rule Mining with a Relative Interestingness Measure*, Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2000, 86–97.
[34] J. Han and M. Kamber, *Data Mining: Concept and Techniques*, Morgan Kaufmann Publishers, 2001.
[35] T. Inoue and S. Abe, *Fuzzy Support Vector Machines for Pattern Classification*, Proceedings of IJCNN, 2001, 1449–1455.
[36] J.J. Jung and G. Jo, *Semantic Outlier Analysis for Sessionizing Web Logs*, Proceedings of the 14th European Conference on Machine Learning/7th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), Cavtat – Dubrovnik, 2004, 13–25.
[37] T. Johnson, I. Kwok and R. Ng, *Fast Computation of 2-D depth Contours*, Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1998, 224–228.
[38] W. Jin, A. K.-H. Tung and J. Han, *Mining Top-n Local Outliers in Large Databases*, Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD, San Francisco, California, USA, 2001, 293–298.
[39] T. Jussi, *Outliers in Non-linear Time Series Econometrics*, PhD Dissertation, University of Turku, Department of Economics, FIN-20014 Turku, Finland, June 2001.
[40] S. Jaroszewicz T. Scheffer, *Fast Discovery of Unexpected Patterns in Data, Relative to a Bayesian Network*, Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, 2005, 118–127.
[41] Y. Kou, C.-T. Lu, S. Sirwongwattana and Y-P. Huang, *Survey of Fraud Detection Techniques Networking*, (Vol. 2), IEEE International Conference on Sensing and Control, 2004, 749–754.
[42] E.M. Knorr and R.T. Ng, *A Unified Notion of Outliers: Properties and Computation*, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1997, 219–222.
[43] E.M. Knorr and R.T. Ng, *Algorithms for Mining Distance-Based Outliers in Large Dataset*, Proceedings of the 24th VLDB International Conference, New York, USA, 1998, 392–403.
[44] E.M. Knorr and R.T. Ng, *Finding Intentional Knowledge of Distance-Based Outliers*, Proceedings of the 25th International Conference on Very Large Databases (VLDB), 1999, 392–403.
[45] E.M. Knorr, R.T. Ng and V. Tucacov, Distance-based outliers: Algorithms and applications, *The VLDB Journal* **8**(3–4) (2000), 237–253.
[46] P. S. Keila and D.B. Skillicorn, Detecting Unusual and Deceptive Communication in Email, Technical Report, School of Computing, Queens University, ISSN-0836-0227-2005-498, 2005.
[47] S. Lin and E.D. Brown, An Outlier-based Data Association Method for Linking Criminal Incidents, Technical Report, Department of Systems Engineering, University of Virginia, SIE020010, 2000.
[48] B. Liu, W.Hsu, L. Mun and H. Lee, Finding interesting patterns using user expectations, *IEEE Transactions on Knowledge and Data Engineering* **11**(6) (1999), 817–832.
[49] B. Liu, W. Hsu, L. Mun and H. Lee, *Pruning and Summarizsing the Discovered Associations*, Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999, 125–134.
[50] B. Liu, Y. Ma and P.S. Yu, *Discovering Unexpected Information from Your Competitors' Web Sites*, Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, 2001, 144–153.
[51] C.R. Miller and A.B. Myres, *Outlier Finding: Focusing User Attention on Possible Errors*, Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology, 2001, 81–90.
[52] S.L, Miller, W.M. Miller and P.J. Mcwhorter, Extrema1 dynamics: A unifying physical explanation of fractals, L/F noise, and activated processes, *Journal of Applied. Physics* **13**(6) (1993), 2617–2628.
[53] F. Provost and J. Aronis, Scaling up inductive learning with massive parallelism, *Machine Learning* **23**(1) (1996), 33–46.
[54] M.I. Petrovskiy, Outlier detection algorithms in data mining systems, *Programming and Computer Software* **29**(4) (2003), 228–237.
[55] C. Piatetsky-Shapiro and C.J. Mathcus, *The Interestingness of Deviations*, Proceedings of AAAI Workshop on Knowledge Discovery in Data Mining Databases, 1994, 25–36.
[56] B. Padmanabhan and A. Tuzhilin, *A Belief-Driven Method for Discovering Unexpected Patterns*, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1998, 94–100.
[57] B. Padmanabhan and A. Tuzhilin, Unexpectedness as a measure of interestingness in knowledge discovery, *Decision Support Systems* **27**(3) (1999), 303–318.
[58] B. Padmanabhan and A. Tuzhilin, *Small is Beautiful: Discovering the Minimal Set of Unexpected Patterns*, Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000, 54–63.
[59] P. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection*, (3rd Edition), John Willey & Sons.

[60] S.J. Roberts, Novelty detection using extreme value statistics, IEE Proceedings on Vision, *Image and Signal Processing* **146**(3) (1999), 124–129.
[61] I. Ruts and P. Rousseuw, Computing depth contours of bivariate points cloud, *Computational Statistics and Data Analysis* **23** (1996), 153–16.
[62] S. Ramaswamy, R. Rastogi and K. Shim, *Efficient Algorithms for Mining Outliers from Large Data Set*, Proceedings of the ACM SIGMOD International Conference, USA, 2000, 427–438.
[63] G. Sheikholeslami, S. Chatterjee and A. Zhang, *WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Database*, Proceedings of the International Conference on Very Large Databases. New York, USA, 1998, 428–439.
[64] P. Smyth and R.M. Goodman, An information theoretic approach to rule induction from databases, *IEEE Transactions on Knowledge and Data Engineering* **4** (1992), 301–316.
[65] D.B. Skillicorn, *Beyond Keyword Filtering for Messages and Conversation Detection*, IEEE International Conference on Intelligence and Security Informatics (ISI), Atlanta, GA, USA, 2005, 231–253.
[66] S. Shekhar, C. Lu and P. Zhang, *Detecting Graph-Based Spatial Outliers: Algorithms and Applications*, Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001, 371–376.
[67] E. Suzuki, *Discovering Interesting Exception Rules with Rule Pair*, Proceedings of the Workshop on Advances in Inductive Rule Learning (with PKDD) 2004, 163–178.
[68] A. Sun, E. Lim and W. Ng, *Web Classification Using Support Vector Machine*, Proceedings of the 4th ACM-WIDM International Workshop on Web Information and Data Management, Virginia, USA, 2002, 96–99.
[69] E. Suzuki and M. Shimura, *Exceptional Knowledge Discovery in Databases Based on Information Theory*, Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, AAAI Press, California, 1996, 275–278.
[70] A. Silberschatz and A. Tuzhilin, *On Subjective Measures of Interestingness in Knowledge Discovery*, Proceedings of the First International Conference on Knowledge Discovery and Data Mining, 1995, 275–281.
[71] A. Silberschatz and A. Tuzhilin, What makes patterns interesting in knowledge discovery systems, *IEEE Transactions on Knowledge and Data Engineering* **8**(6) (1996), 970–974.
[72] J. Tang, Z. Chen, A. Fu and D. Cheung, *Enhancing Effectiveness of Outlier Detections for Low Density Patterns*, Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Taipei, Taiwan, 2002, 535–548.
[73] J.W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, 1977.
[74] P.H. Torr and D.W. Murray, Outlier detection and motion segmentation, in: *Journal of International Society for Optical Engineering* (*SPIE*), (Vo. 2059), Paul S. Schenker, ed., 1993, pp. 432–443.
[75] http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data, December 2004.
[76] W. Wang, J. Yang and R. Muntz, *STING: A Statistical Information Grid Approach to Spatial Data Mining*, Proceedings of the 23rd VLDB International Conference, Greece, 1997, 186–195.
[77] K. Yamanish, J. Takeuchi, *A Unifying Framework for Detecting Outliers and Change Points from Non-Stationary Time Series Data*, Proceedings of the 8th ACM SIGKDD International Conference, Canada, 2002, 676–681.
[78] D. Zhang and S.W. Lee, *Question Classification Using Support Vector Machines*, Proceedings of the 26th Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval, Toronto, Canada, 2003, 26–32.
[79] J. Zhao, C. Lu and Y. Kou, *Detecting Region Outliers in Meteorological Data*, Proceedings of the 11th ACM International Symposium on Advances in Geographic Information Systems, 2003, 49–55.
[80] T. Zhang, R. Ramakrishnan and M. Linvy, *BIRCH: An Efficient Data Clustering Method for Very Large Databases*, Proceedings of ACM SIGMOD International Conference on Management of Data, ACM Press, New York, 1996, 103–114.