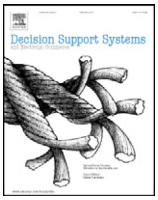
Decision Support Systems 75 (2015) 63–75

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss



Algorithm for the detection of outliers based on the theory of rough sets

Francisco Maciá-Pérez [a], Jose Vicente Berna-Martinez [a,*], Alberto Fernández Oliva [b], Miguel Alfonso Abreu Ortega [b]

[a] Department of Computer Technology, University of Alicante, Carretera San Vicente del Raspeig s/n - 03690 San Vicente del Raspeig, Alicante, Spain
[b] Department of Computer Science, School of Mathematics and Computer Science, University of Havana, Cuba

ARTICLE INFO

Article history:
Received 12 March 2014
Received in revised form 12 March 2015
Accepted 3 May 2015
Available online 12 May 2015

Keywords:
Knowledge discovery
Detection of outliers
Rough set theory

ABSTRACT

Outliers are objects that show abnormal behavior with respect to their context or that have unexpected values in some of their parameters. In decision-making processes, information quality is of the utmost importance. In specific applications, an outlying data element may represent an important deviation in a production process or a damaged sensor. Therefore, the ability to detect these elements could make the difference between making a correct and an incorrect decision. This task is complicated by the large sizes of typical databases. Due to their importance in search processes in large volumes of data, researchers pay special attention to the development of efficient outlier detection techniques. This article presents a computationally efficient algorithm for the detection of outliers in large volumes of information. This proposal is based on an extension of the mathematical framework upon which the basic theory of detection of outliers, founded on Rough Set Theory, has been constructed. From this starting point, current problems are analyzed; a detection method is proposed, along with a computational algorithm that allows the performance of outlier detection tasks with an almost-linear complexity. To illustrate its viability, the results of the application of the outlier-detection algorithm to the concrete example of a large database are presented.

1. Introduction

Decision support systems are computer-based programs that assist decision makers in effective and efficient decision-making. The proper functioning of these systems requires large amounts of precise, high-quality data. However, if the data contain abnormal, unrealistic or simply erroneous elements, it may misguide the decision-making process, thereby leading to incorrect results. These abnormal elements must be detected, isolated and analyzed to check whether they have any real meaning or simply represent a monitoring error. Currently, data sets in the real world and their environments present a wide range of difficulties that limit the efficiency of some existing detection methods. One of the most noteworthy problems is that data sets can be very large and dynamic, imposing the need for efficient algorithms with regard to time complexity. Recent investigations related to knowledge discovery in databases (KDDs) have paid special attention to issues regarding the detection of outliers, which become more serious with the large volumes of information stored in today's databases [1,2]. If, in general, KDD-data mining (KDD-DM) processes are directed toward the discovery of representative behavioral patterns, the detection of outliers takes advantage of the high marginality of these objects (marginality refers to how much or little different which is an element from the rest), and they are detected by measuring their degree of deviation with respect to the aforementioned patterns. From the perspective of KDD-DM, outlier detection can be viewed in two different ways: outliers can be considered undesirable objects that must be treated or eliminated at the stage of preparation of the data, as their presence in the set may interfere with the efficient detection of trustworthy patterns [3]; or they can be considered as objects that should be identified for their implicit relevance to the processing of the data [2]. In the latter case, they must not be eliminated from the data set, because for some applications, outliers are more representative and interesting than common events from the point of view of information discovery. Therefore, outlier detection is also a process of information (knowledge) discovery and is of great utility for the analysis and interpretation of data.

The range of applicability of outlier detection is very wide and diverse, and applications can be found in widely varied scenarios. In wireless networks, outlier allows for the detection of atypical readings and their subsequent correction [4]. By monitoring activities of various kinds, such as the activity of a mobile phone or an online store, it is possible to detect suspicious or unlawful activities [5]. In the study of DNA irregularities, outlier detection may lead to the discovery of genetic alterations that result in disease and structural defects [6]. In the automated control of assembly lines, it allows the detection of production defects [7]. In pharmaceutical research, it aids in the identification of new molecular structures [8]. This diversity in the range of application is one of the justifying motives for the wide variety of existing methods

* Corresponding author. Tel.: +34 965 90 3400x1307; fax: +34 96 590 9643.
E-mail addresses: pmacia@dtic.ua.es (F. Maciá-Pérez), jvberna@dtic.ua.es (J.V. Berna-Martinez), afdez@matcom.uh.cu (A. Fernández Oliva), miguel87@lab.matcom.uh.cu (M.A. Abreu Ortega).

http://dx.doi.org/10.1016/j.dss.2015.05.002
0167-9236/

of outlier detection. The data in each situation possesses a distinct nature and definition space, and therefore, the detection methods must adjust to the data types and contexts where they will be applied [9]. Therefore, the search for efficient methods that may be utilized in any situation, which are more flexible, adaptable and scalable, is a problem of great interest.

Some of the techniques being applied efficiently in KDD-DM processes are related to Rough Set Theory [10,11], which is advantageous due to its flexibility and adaptability to different scenarios. This adaptability is demonstrated by the variety of related works found in the literature, including the process of evaluation of complex information systems [12], learning in neural networks [13], analysis of reconfigurable manufacturing systems (RMS) [14], solutions to problems in the field of investment [15], application to the grid scheduling process [16], adjudication of bank credits [17], image processing [18], and the evaluation of business innovation capabilities [19]. Thus, the ability of these techniques to model a wide range of real-world situations, their efficiency in the resolution of various types of problems, and their wide range of applicability have been made manifest.

The use of rough sets (RSs) extends within the KDD-MD and is also beginning to be utilized as the foundation for the characterization and detection of outliers. This approach is a novel point of view with great potential that shows promise for the construction of efficient algorithms [20,21], capable of detecting outliers with a high degree of marginality. However, these detection schemes have as a disadvantage the inconvenience of using the concept of non-redundant exceptional sets to classify the most contradictory elements of a data set, from which the outliers are obtained. The weak point of this scheme is that identifying such sets requires the identification of a power set, which leads to a problem of exponential time complexity ($\Omega(2^n)$, n being the cardinality of the data set). In today's world, where data volumes are of great size, this problem makes such schemes unfeasible from a computational point of view, even though they may be formally sound in mathematical rigorous terms.

The present article analyzes the problem of exponential temporal complexity exhibited by current algorithms based on rough set theory [20]. We propose an expansion of the existing mathematical framework, which allows for the creation of a method of outlier detection based on rough sets that is in fact computationally viable, with a corresponding algorithm of almost linear time complexity. Furthermore, a runtime study of the use of the proposed algorithm applied to a realistic data set is presented, showing that, indeed, it shows in practice the behavior that is described mathematically.

The remaining sections of the article have been organized as follows: Section 2 summarizes the current state of the technique, along with some background for this research work, relevant aspects of RS theory, and its main inconveniences. In Section 3, we discuss the foundations of rough set theory and present a simple example to clarify its inner workings. In Section 4, the expansion of the mathematical framework is developed, which will subsequently allow the proposal of the computational algorithm. In Section 5, an algorithm is proposed for the detection of outliers based on the basic model of rough sets, emphasizing specific aspects of its implementation. In Section 6, the different tests that were performed with the proposed algorithm on a real data set are shown, corroborating the theoretical results. Finally, in Section 7, the main conclusions of this work are presented, along with the main lines of future research. In the Appendix at the end of this work, the theoretical framework proposed in Section 4 is explicitly demonstrated.

2. Background

Generally speaking, in data mining, the detection of outliers allows for the identification of unexpected input in a database and, on these grounds, the determination of various types of errors, data usage fraud, the existence of valid but atypical values, and many other features of interest. Therefore, its applications are highly diverse, for example, the detection of intruders in computer networks [22]; the data mining of manuscripts in the context of a project for the digitalization of the cultural and scientific heritage of Bulgaria [23]; applications in medical diagnosis, where outlier detection can aid in the diagnosis of a given pathology [24]; the detection of outliers in climate studies related to the temperature ranges in different world cities [25]; the analysis and processing of population data for the US Census Bureau's Income report [26–28]; the detection of traffic risks based on which accident prevention measures can be taken [29]; outlier detection techniques in data sets of credit card usage information, employed to detect their misuse [30]; outlier detection used for the adequate classification of crystals based on chemical and physical tests [30]; in the context of sports, outlier detection used to monitor the performance of NBA players in the USA [31]; and in video surveillance, outlier detection allows guaranteed safety in public areas (video/image data mining) [32]. As can be appreciated, outlier detection is a subject of applicability as vast as the existing types of databases.

Due to the great number of scenarios and data types, different approaches to the problem of outlier detection have arisen, and above all, proposals that address large data volumes have begun to gain importance. This problem was treated first in the field of statistics: statistical models are generally appropriate for the processing of data sets with quantitative, real, continuous values, or at least qualitative data with ordinal values. Nowadays there is an ever-increasing need for the processing of categorical (non-ordinal) data. This requirement considerably limits the applicability of statistical methods. Another deficiency of statistical methods is their limited functionality in high-dimensionality (multivariable) spaces, where it is generally exceedingly difficult to find adequate models [33].

There are methods based on non-parametric approximations, among which distance-based outlier detection methods can be found. One of the most widely utilized is the method of k-nearest neighbors (K-NN) [34,35]. There are different approaches to the K-NN algorithm, but all use a metric that is appropriate for the calculation of distances between neighbors, such as the Euclidean distance or the Mahalanobis distance. There are also proposals that optimize the basic K-NN algorithm [36].

In general, it can be said that there are numerous techniques for the detection of outliers in which algorithms of different kinds are combined [20,37]. Among the most outstanding methods, some categories can be identified, such as methods based on distributions [38], depth [39], distances [34,36], densities [40], clusters [41], or support vectors [42]. Most recent research works address various detection methods based on artificial intelligence techniques, fundamentally, techniques related to machine learning [43]. In [44], we find a very complete compilation of the most outstanding outlier detection methods. Most of the distance-based methods are of at least quadratic of time completion order with respect to the number of elements in the data set, which may be unacceptable if the data set is very large or dynamic. On a different front, statistical methods essentially center on the detection of outliers among single-variable data. They require a priori knowledge of the data distribution. In these cases, the user must model the data utilizing a statistical distribution, and the outliers are determined depending on how they appear in relation to the postulated model. The main problem with this approach lies in the number of possible situations and on the possibility that the user may lack sufficient knowledge of the data distribution.

Considering that no universally applicable outlier detection approach is available and that researchers must focus their efforts on the selection of an acceptable method for their specific data set, this subject still poses a very open problem. A consequence is the continued appearance of new models and new methods based on a diversity of schemes and approaches to the problem at hand. One of these new propositions is the application of Rough Set Theory to outlier detection, where