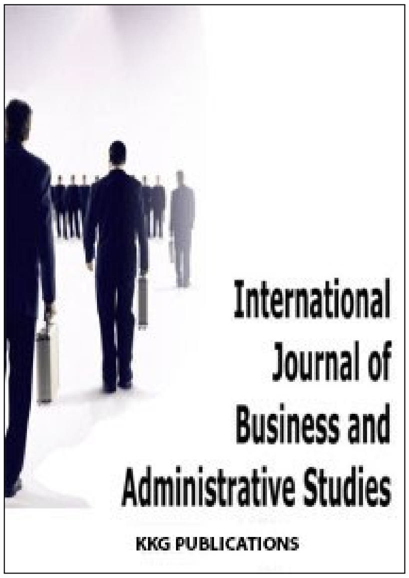


Key Knowledge Generation
Publication details, including instructions for author and subscription information:
http://kkgpublications.com/business/

A Novel Data Mining Study to Spot Anomalies in Organizations: A Human Resources Management Case

GOKHAN SILAHTAROGLU [1], PELIN VARDARLIER [2]

Istanbul Medipol University, Turkey

Published online: 10 August 2016

To cite this article: Silahtaroglu, G., & Vardarler, P. (2016). A novel data mining study to spot anomalies in organizations: A human resources management case. *International Journal of Business and Administrative Studies, 2*(4), 110-116.
DOI: https://dx.doi.org/10.20469/ijbas.2.10001-4

To link to this article: http://kkgpublications.com/wp-content/uploads/2016/2/Volume2/IJBAS-10001-4.pdf

International Journal of Business and Administrative Studies
2016, 2(4), 110-116
IJBAS

A NOVEL DATA MINING STUDY TO SPOT ANOMALIES IN ORGANIZATIONS: A HUMAN RESOURCES MANAGEMENT CASE

GOKHAN SILAHTAROGLU [1]*, PELIN VARDARLIER [2]

Istanbul Medipol University, Turkey

Keywords:
Data Mining
Anomaly Detection
Employee Behavior
Organizational Behavior

Received: 30 January 2016
Accepted: 28 March 2016
Published: 10 August 2016

Abstract. Organizational behaviour is one of the most important assets of an organization; it may increase business value and profitability. Although it comes to life itself with the sector's contribution, leadership, environment, market conditions and competition, unemployment etc., there must be some ways to lead or direct an organizational behaviour. However, it is not practical to monitor and control organizational behaviour in detail with bare eyes. Technology-wise, we are not away from building a system to learn, understand and monitor the organizational behaviour and give alarms or signals to top management when the system perceives an extraordinary condition. This is doable with machine learning algorithms which is a model of Artificial Intelligence (AI). In the literature, there are plenty of machine learning algorithms that have been proved that they work or learn very well. Both supervised and unsupervised algorithms may be used to learn organizational behaviour and detect anomalies in the behaviour. In this paper, we propose a novel system to build a data warehouse with the data available in an organization and design it for learning the organizational behavior by machine, i.e. computers. The system we propose may be used to spot daily, weekly or monthly changes in organizational behaviour. These changes may have positive or negative effects on the performance of organizations, so this system may also be used as a decision support system for top management. When necessary feedback is given to the system, it may also develop or learn new features to interpret the effect of the change on the organization's overall performance. The system may also be used to compare the changes in organizational behaviour every year. Nevertheless, it is a useful tool to track, learn and monitor the interactions of each employee in the system. In the study, human resources management has been used as a sample model to detail the proposed system.

INTRODUCTION

Organizational behavior is a field of study that researches the effects of staff and organizational structure on behavior of the organization. The main purpose of this field is to increase organizational effectiveness. From the definition above, we can say that the study of organizational behavior has three main elements; one element is that organizational behavior is an analytical study of individuals and groups; secondly, the impact of organizational structure over human behavior and the third, the application of knowledge to achieve organizational effectiveness. These objectives and targets are achieved with various methods. Therefore it is clear that organizational behavior is an important factor to boost up the organizational brand value, profitability, effectiveness etc. of a business or organization. Since manipulating organizational behavior may help the business gain, then any abrupt changes in the organizational behavior may signal something. It may be positive or negative. Thus, monitoring and learning organizational behavior is crucial for a business. This monitoring and therefore learning the natural activities may be realized with modern machine learning algorithms within the understanding of data mining concept.

Data mining has been popular with the increase of databases and has been used widely for different purposes and cases. It has three accepted models: Classification, clustering and association rule discovery. Classification employs supervised machine learning algorithms such as C4.5, SPRINT, Artificial Neural Networks and so on. In classification practice, a certain class must be chosen by the researcher in order to form a learning model to predict unknown classes of some other data. In this process, data available for research are divided into two parts: Data for learning or discovering hidden rules and data to validate the rules discovered by the model or algorithm(s). Association rule discovery is another model of data mining. In literature, it is also named as link analysis. This model of data mining uses pattern recognition algorithms such as APRIORI and GRI. This model searches repeated and meaningful patterns as item set or sequential order. It produces and learns rules like if item A exists and item B exits then item C also exists. If the model is set up to produce sequential patterns it will generate rules like if item A exits and item B exits then in X unit time item C may exist.

*Corresponding author: Gokhan Silahtaroglu
†Email: gsilahtaroglu@medipol.edu.tr

Clustering is the other model of data mining which exercises unsupervised machine learning algorithms. In this model, there is no class to be predetermined by users. In other words, there is no label for any record such as good, bad, high, low, normal etc. Clustering algorithms simply divide the given data into sub clusters. Theoretically unsupervised algorithms decide the number of sub-clusters and also the number of members in each cluster. These algorithms while doing this partitioning process also have a chance to spot outliers in the data set, at the same time. Outliers are records in a data-set that do not belong to any cluster; neither can they form another cluster by themselves.

In this paper, we propose a model to be used in any business sector to detect employee behavior anomaly together with organizational behavioral anomalies. In the second part of the paper we present recent studies on anomaly detection in various cases, sectors and application areas. In the third part of the paper, we present a model to be used within an Enterprise Resource Planning (ERP) or any office automation system.

ANOMALY DETECTION IN LITERATURE

As it is mentioned earlier clustering is used to partition the given data set into sub-clusters. In this way outliers which do not belong to any clusters are also detected. If a record belongs to a sub-cluster we may assume that it is quite common to encounter that record any time. However, it is not usual to encounter an outlier very often. Because outliers are rare items in the data sets, but this is not to say that an outlier is a faulty record that entered the data set accidentally. Simply, outliers are rare. If a machine is working properly for thousands of hours its data recorded will form some cluster(s). When the machine started to work inefficiently and finally produced wrong output and collapsed, it would give out different set of recordings (logs). Probably, these recordings will not fit to any clusters formed when the machine was working properly. In this case, the data recorded will be outlier since they will not fit to any clusters formed beforehand. In literature there are plenty of studies and applications that used data mining and clustering for anomaly detection or any similar purposes. For example, in a recent study Pacacios et al. proposed a model to detect situations when aero engines are not working in normal conditions (Palacios, Martinez, Sanchez & Couso, 2015). In the study they used numerical pattern recognition algorithms to obtain fuzzy rules from Engine Health Monitoring (EHM) data. They have managed to identify deterioration levels of an aero engine at any given time. For this purpose, they used an adaptation of PrefixSpan algorithm (Palacios et al., 2015). Simple defects at moving parts in a machinery would cost a lot if necessary precautions were not taken on time. Stressing the necessity of this, Purarjomandlangrudi presented an approach to detect anomaly in wind turbine bearings (Purarjomandlangrudi, Ghapanchi, & Esmalifalak, 2014). They used kurtosis and Non-Gaussianity Score (NGS) parameters to detect anomaly in advance and tested the approach against Support Vector Machines (SVM) using real data. A similar study was held in 2010 to detect rolling bearing fault by Al-Raheem and Kareem (2010). In these studies artificial neural networks and wavelet transforms were used to diagnose fault in machines. Demetgul, Yildiz, Taskin, Tansel and Yazicioglu, (2014) conducted an experimental research to diagnose fault at pneumatic systems of the material handling systems. They collected data from the pneumatic systems that were operating at both normal and faulty conditions. They used Diffusion Map (DM) (Coifman et al., 2005), (Lafon & Lee, 2006), Local Linear Embedding (LLE) (Roweis & Saul, 2000) and Auto Encoder (AE) (Van der Maaten, 2007) algorithms for feature extraction and GustafsonKessel (GK) (Sarbu, Zehl & Einax, 2007) and k-medoids (Zhang & Couloigner, 2005) algorithms for classification. In this way they showed that faulty conditions of pneumatic systems can be predicted beforehand. Apart from anomaly and fault detection at operating machines, data mining has also been used for other systems to detect unusual and unwanted patterns. One of these systems is network systems. Cerroni, Moro, Pasolini, and Ramilli, (2015) used data mining clustering algorithm (k-Means) to detect the pattern of network attacks. For this purpose they inspected network traffic packets for anomalies, in other words network attacks. In order to extract the universal feature of attackers, Xia et al., (2015) proposed a novel anomaly detection approach. They used a clustering or segmentation technique to confirm the size of the time interval dynamically to group consecutive attack ratings. In another research, synchronization anomalies in parallel programs have been detected by analyzing data and control dependencies. In this way it is shown that some information may be gathered for bug fixing in computer programs (Jannesari, 2015). Similar studies were held in the field of medicine, marketing, banking etc. As it is seen different clustering or segmentation algorithms and techniques have been used successfully to spot anomalies in data sets.

Outlier Detection Methods and Algorithms

In the literature, there are five basic methods to learn and detect anomalies (Hemalatha, Vaidehi & Lakshmi, 2015). Each method entails various algorithms and techniques. The researcher or the expert should decide on which method and algorithm to use. For one case one method or algorithm may work well as for another case a different method and algorithm may.

Clustering

Clustering algorithms are designed to find sub clusters within a given data set so that each sub cluster or segment members will be more similar to their own cluster members than any other non-members. In the Figure-1 there is a two dimensional data set. A proper algorithm is to divide the data set into three clusters as in Figure-2. Those which do not belong to any of these three clusters will be considered as outliers. Clustering algorithms employ distance measures. The most common distance measures which are used within clustering algorithms are Euclid and Manhattan distance measures (Knox & Ng 1998).

FIGURE 1
A Two Dimensional Data Set Representation

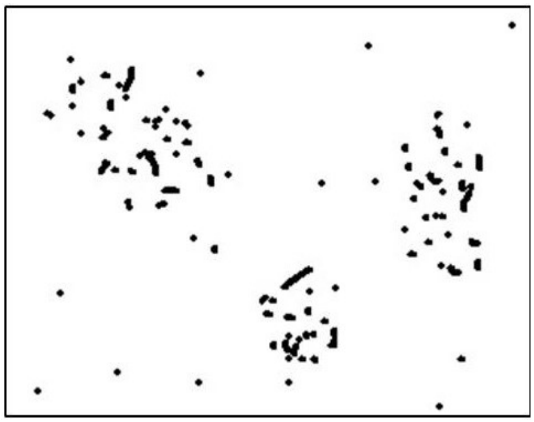
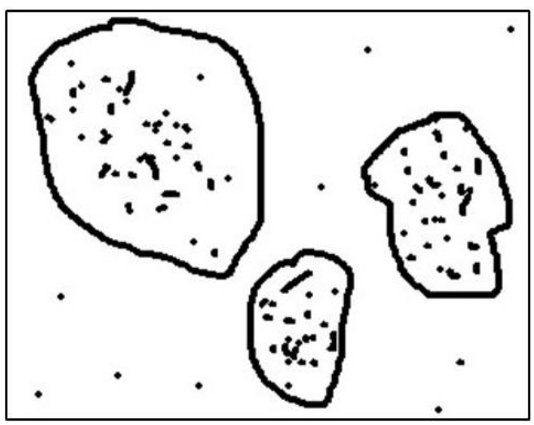


FIGURE 2
Data Set has been Divided into Three Clusters



Distance-Based Methods

Distance-based methods or algorithms are similar to clustering as both of them use one or another kind of distance measure. However, distance-based methods usually depend on one or more parameters like minimum / maximum distance between clusters or records, minimum / maximum number of records in each cluster. Knorr, Ng and Tucakov (2000), Angiulli and Pizzuti (2002), Todeschini, Ballabio, Consonni, Sahigara and Filzmoser, (2013), Wang, Xiao, Yu and Yang (2009) are some of the distance-based algorithms to detect outliers. All of these algorithms may be exploited as anomaly detection.

Frequent and Infrequent Pattern Mining (Rare Item-Set Mining)

Item-set mining is used to extract pattern that may be exploited for different business purposes. However, rare item-set mining is somehow different. It is to extract rare items in a dataset as its name suggests so, it is also used for outlier detection, or to be more specific with our subject, the algorithms in this category may also be used for anomaly detection. There are plenty of algorithms and case studies about rare item-set mining and anomaly detection (Hemalatha et al. 2015).

Density-Based Methods

Distance measures may not work in some situations as there are various patterns in the data set. Some patterns may cause distance-based algorithms to fail. Density-based algorithms search clusters that form a relative density within the whole data set. These algorithms start with a single record and check all around density to detect clusters. Records that do not exist within proximity of defined density are considered to be anomaly in that data set. On the other hand, some the algorithms take the whole data set as a single cluster at first, and check the gaps and dense areas in the data set which form clusters and detect outliers at the same time (Ester, Kriegel, Sander & Xu, 1996; Breunig, Kriegel, Ng & Sander, 2000).

Artificial Neural Networks

Artificial Neural Network (ANN) is a model of knowledge discovery inspired by biological neural system. ANN consists of artificial cells connected with each other in a predesigned order. Each connection carries a w parameter to be calculated within the system. Those weights are the multipliers of real data entries and generated values in the system (Yegnanarayana, 2009). With methods like back propagation, feed-forward or delta rule etc. errors are minimized to reach the best solution as of clusters. After clustering is completed, each record or data entry is labeled with the name of the cluster it belongs to. ANN also generates accuracy level of each cluster member to be in that cluster. Some records may be eliminated and labeled as outliers using this accuracy percentage value(s). An artificial neural network has input layers, output layers and multiple hidden layers between them. Each layer contains at least one node to connect both layers and data records with their outputs (Hagan, Demuth & Beale, 1996). Each node is calculated with the weight of the connection and the value coming from the data set. After that the node makes another calculation using a proper activation function. Some of the activation functions to be used are sigmoid, log-sigmoid, step and softmax functions (Yegnanarayana, 2009; Haykin, Haykin, Haykin & Haykin, 2009).

OUR WORK AND MODEL

Today, in any organization, a software system like Enterprise Resource Planning (ERP), Office Automation System (OAS) or a simple mailing system exists. Every day, all interactions, memorandums, meeting minutes, database access logs etc. are stored. Nevertheless, very little part of these data sources is converted into reports or information which will be profitable to the organization. These data may be used for anomaly detection processes throughout the organization though.

Under normal conditions, individuals or people draw a pattern when they commute to and fro, eat meals, spend their weekends etc. Either this way or that way, each of us has a life or activity pattern formed with our natural deeds. Again, every employee working in a company forms a general but steady working pattern when s/he performs tasks in the business. When we think of the whole organization, all these individual activities will be the behavior of the business itself. Indeed, it will be a certain and repeating pattern that can be recognized and predicted when necessary software programs enhanced with proper algorithms are used. This is possible with tracking the activities stored in databases with appropriate data mining models and machine learning algorithms.

Let's clarify this with a real life procedure that can take place in typical business; Personnel Recruitment:

A typical personnel recruitment process requires following steps:

- A form is filled out by the department which needs new staff.
- This need is confirmed by related member of the top management and passed to the human resources department.
- Human resources department is given a certain period of time to complete the necessary procedures and fill the vacancy.
- After that a series of interviews are held with candidates taken from a Resume Database or a job advertisement.
- Candidate testing is done.
- References are controlled.
- Qualified candidates are directed to the department which has the vacancy.
- The department either accepts one of the candidates or rejects all of them. In this case, all procedure loops or vacant position and all the procedures are cancelled.
- If one of the candidates is found qualified enough to fill the vacant positions, new procedures will be initiated: Job offer, registration of the new employee, orientation process and finally closing the position.

Vardalier et al. presents it with an Axiomatic Design as it is in Table 1.

All of the above mentioned activities are put in computer through proper software in organizations or firms. In this study, we propose that all these activities, transactions, records whatever you call it, is a pattern that repeats itself and this pattern is predictable. A machine learning algorithm may learn not only this pattern but also all other patterns in a business and mav predict next steps with a certain error. This is what a typical machine learning algorithm does. If the algorithm has learnt it well it will make its predictions within a certain error range under normal business conditions. If the algorithm fails to predict next step(s) or the error range goes up, then we may conclude that there is an anomaly in the system. For example, if the overall average of Human Resources Department (HRD) is to hold five interviews in two days under a certain condition and the department performs only one (way below average) or ten (double of the average), it may be taken as an extraordinary situation or anomaly. This means that something has been changed positively or negatively in that part of the business during those days.

TABLE 1
Recruitment Model by Axiomatic Design (Vardarlier, Vural & Birgun, 2014)

FR0: Increases the Efficiency of the Recruitment Process	DP0: Standardized Recruitment Model
FR1: Determine the Open Position	DP1: Work Analysis, Job Study, Job Description, Norm Staff
FR1.1: Interrogate the Open Position	DP1.1: Norm Staff
FR1.2: Define the Open Position	DP1.2: Job Description
FR2: Determine the Sources From Where the Candidates Will Be Supplied	DP2: Evaluation of Candidates Pool, Evaluation of Personnel Structure
FR2.1: Check Internal Sources	DP2.1: Organization Chart, Career Planning, Promotion Opportunities and Rotation
FR2.2: Analyze External Sources	DP2.2: Current Resume Pool, Transfer Opportunities, Consulting Services, Market Analysis
FR3: Do the Bulletin and Announcement Plan	DP3: Bulletin and Announcement Process, Map of Competence
FR4: Make the Bulletins and the Announcement on Appropriate Platforms	DP4: New Paper, Social Media, Online Announcement Portals, Internal Announcements and Consultancy Firms
FR5: Determine the Interview and Test Processes	DP5: Appropriate Evaluation Methods Depending on the Requirements of the Position (Inventory, Interview, Competence Based Evaluation, Group Interview)
FR6: Complete the Preliminary Evaluation Process	DP6: Candidate Pool, Competence Map, Recruitment Process, Job Description
FR7: Realize the Test and Inventory Sourced Evaluation.	DP7: Appropriate Test and Inventory Techniques
FR8: Establish a Candidate Short List	DP8: Candidate Selection
FR9: Evaluate the Short List	DP9: Test, Inventory and Techniques, Interview Process
FR9.1: Complete Technical Assessment	DP9.1: Assessment Center, Competency Map
FR9.2: Perform the Job Interview Process	DP9.2: Individual Interview, Group Interview, Role Play, Case Study
FR9.3: Check the Background	DP9.3: Reference Check, Application Documents
FR10: Final Decision	DP10: Results of the Interview, Scores, Reference Control
FR10.1: Select the Appropriate One of the Current Candidates	DP10.1: Selection
FR10.1.1: Create the Job Offer	DP10.1.1: Wage Package
FR10.1.2: Start the Orientation Process	DP10.1.2: Orientation, Job Training
FR10.1.3: Close the Position	DP10.1.3: Feedback
FR10.2: If Candidates Don't Provide Sufficient Criteria, Start All the Process Again.	DP10.2: Recruitment Process, Source Planning, Interview, Evaluation, Decision

However, a special data warehouse is needed for this learning and anomaly detection process. This data warehouse should be made from following data sources:

- Transactions of Accounting Department.
- Internal Mails
- Mails coming from outside the company
- Any kind of forms filled out in the business
- Typical business transactions, accomplishments or activities like orders, cancelations, purchases, job interviews etc.
- CVs collected for future recruitment or being processed.
- Voice calls converted into text form if applicable.
- Server activities or log files.
- Printer and any other device usage activity logs.
- Database access logs.
- Activities like holiday or take leave planning or their records.

The data warehouse to be used for organizational behavior anomaly is not limited to the data sources mentioned above. However it is clear that this data warehouse will be made up of data in different forms like text, numbers, images, voice etc. So, all of them must be put into a single form to be analyzed or learnt by machine learning algorithms to detect future anomalies in the organization. Here we had better stressed that each record should have a suitable timestamp. The learning and anomaly detection will be done all around the timestamp used in the data warehouse. Hence, the type of the timestamp is also very important for the performance and success of the proposed system.

DISCUSSION AND CONCLUSION

In this study, we propose a system and a data warehouse to monitor and learn typical organizational activities or behaviors. The system will be a data mining model enriched and augmented with machine learning algorithms. Both supervised and unsupervised algorithms may be tailored and adapted to the system. The system we propose will be an anomaly detection system. After learning typical activities in the organization the model will start to monitor the organization and at the same time it will make predictions about the next step to be taken by the organization. Consequently, none of these predictions will be 100% correct; however the system will also learn its own error rate and will try to sustain a small error. Finally, the system will decide on its own error range and learning will be completed. After that, while the system continues monitoring and adjusting its error, it will also be set up to give signal if there is an abrupt change in the prediction error or if the system encounters an outlier when it is clustering the behaviors. In order to achieve above mentioned success the system needs a data warehouse which is made up of multiple and various data and data sources. These data sources and type of the data may dramatically change from one organization to another. A data mining expert should decide on the type and the structure of the data to be converted into a data warehouse. Time and cost will also be important factor to be taken into account when building the data warehouse.

REFERENCES

Al-Raheem, K. F., & Abdul-Karem, W. (2010). Rolling bearing fault diagnostics using artificial neural networks based on Laplace wavelet analysis. *International Journal of Engineering, Science and Technology, 2*(6), 278-290.

Angiulli, F., & Pizzuti, C. (2002). Fast outlier detection in high dimensional spaces. Paper presented at *PKDD '02 Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, (pp. 15-26).

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *ACM Sigmod Record, 29*(2), 93-104.

Cerroni, W., Moro, G., Pasolini, R., & Ramilli, M. (2015). Decentralized detection of network attacks through P2P data clustering of SNMP data. *Computers & Security, 52*, 1-16.

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., & Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. Paper presented at the *Proceedings of the National Academy of Sciences of the United States of America, 102*(21), 7426-7431.

Demetgul, M., Yildiz, K., Taskin, S., Tansel, I. N., & Yazicioglu, O. (2014). Fault diagnosis on material handling system using feature selection and data mining techniques. *Measurement, 55*, 15-24.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Knowledge Discovery and Data Mining, 96*, (34), 226-231.

Hagan, M. T., Demuth, H. B., & Beale, M. H. (1996). *Neural network design* (pp. 2-14). Boston, Massachusetts, MA: PWS Pub.

Haykin, S. S., Haykin, S. S., Haykin, S. S., & Haykin, S. S. (2009). *Neural networks and learning machines*. Upper Saddle River, New Jersey, NJ: Pearson Education.

Hemalatha, C. S., Vaidehi, V., & Lakshmi, R. (2015). Minimal infrequent pattern based approach for mining outliers in data streams. *Expert Systems with Applications, 42*(4), 1998-2012.

Jannesari, A. (2015). Detection of high-level synchronization anomalies in parallel programs. *International Journal of Parallel Programming, 43*(4), 656-678.

Knorr, E. M., Ng, R. T., & Tucakov, V. (2000). Distance-based outliers: Algorithms and applications. *The International Journal on Very Large Data Bases, 8*(3-4), 237-253.

Knox, E. M., & Ng, R. T. (1998). Algorithms for mining distancebased outliers in large datasets. Paper presented at *Proceedings of the International Conference on Very Large Data Bases* (pp. 392-403).

Lafon, S., & Lee, A. B. (2006). Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 28*(9), 1393-1403.

Palacios, A., Martinez, A., Sanchez, L., & Couso, I. (2015). Sequential pattern mining applied to aeroengine condition monitoring with uncertain health data. *Engineering Applications of Artificial Intelligence, 44*, 10-24.

Purarjomandlangrudi, A., Ghapanchi, A. H., & Esmalifalak, M. (2014). A data mining approach for fault diagnosis: An application of anomaly detection algorithm. *Measurement, 55*, 343-352.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science, 290*(5500), 2323-2326.

Sarbu, C., Zehl, K., & Einax, J. W. (2007). Fuzzy divisive hierarchical clustering of soil data using GustafsonKessel algorithm. *Chemometrics and Intelligent Laboratory Systems, 86*(1), 121-129.

Todeschini, R., Ballabio, D., Consonni, V., Sahigara, F., & Filzmoser, P. (2013). Locally centred Mahalanobis distance: A new distance measure with salient features towards outlier detection. *Analytica Chimica Acta, 787*, 1-9.

Van der Maaten, L. J. P. (2007). *An introduction to dimensionality reduction using matlab*.

Vardarlier, P., Vural, Y., & Birgun, S. (2014). Modelling of the strategic recruitment process by axiomatic design principles. *Procedia-Social and Behavioral Sciences, 150*, 374-383.

Wang, B., Xiao, G., Yu, H., & Yang, X. (2009). Distance-based outlier detection on uncertain data. Paper presented at *Computer and Information Technology, 2009*. CIT'09. *Ninth IEEE International Conference on* (Vol. 1, pp. 293-298). IEEE.

Xia, H., Fang, B., Gao, M., Ma, H., Tang, Y., & Wen, J. (2015). A novel item anomaly detection approach against shilling attacks in collaborative recommendation systems using the dynamic time interval segmentation technique. *Information Sciences, 306*, 150-165.

Yegnanarayana, B. (2009). *Artificial neural networks*. Delhi, India: PHI Learning Pvt. Ltd.

Zhang, Q., & Couloigner, I. (2005). A new and efficient k-medoid algorithm for spatial clustering. In *Computational Science and Its ApplicationsICCSA 2005* (pp. 181-189). Berlin, Germany: Springer Berlin Heidelberg.

–This article does not have any appendix. –