Advanced Data Analytics for IT Auditors

Do you have something to say about this article?
Visit the *Journal* pages of the ISACA® web site *(www.isaca.org/journal)*, find the article and click on the Comments link to share your thoughts.

Data analytics is a must-have capability for the audit function[1] and widely expected to become a big part of its future.[2]

Data analytics is defined as, "the science of examining raw data with the purpose of drawing conclusions about that information..."[3] The definition continues, stating:

> *The science is generally divided into exploratory data analysis (EDA), where new features in the data are discovered, and confirmatory data analysis (CDA), where existing hypotheses are proven true or false... In information technology, the term has a special meaning in the context of IT audits, when the controls for an organization's information systems, operations and processes are examined. Data analysis is used to determine whether the systems in place effectively protect data, operate efficiently and succeed in accomplishing an organization's overall goals.*[4]

Spiros Alexiou, Ph.D., CISA
Is an IT auditor who has been with a large company for eight years. He has more than 20 years of experience in IT systems and data analytics and has written numerous sophisticated computer programs. He can be reached at *spiralexiou@gmail.com.*

Numerous disciplines use simple and advanced data analytics for:

- **Classification**—Identifying good customer/bad customer and fraud/no fraud
- **Clustering**—Identifying groups with similar behavior
- **Association**—Determining that everyone who bought item A also bought item B, and 80 percent of them also bought item C
- **Summarization**—Describing groups with certain characteristics (e.g., executives with average use of company card totals greater than **x** dollars)
- **Link analysis**—Determining connections (e.g., A called B, and B immediately called C, hence, A may be linked to C)
- **Deviation detection**—Identifying transactions significantly different from the average
- **Prediction/estimation**—Predicting trends or growth of a new business
- **Visualization**—Perhaps this is not data analytics proper, but aids in nonautomated human discovery (e.g., charts or medical imaging)

Two Categories of Data Analytics

Data analytics techniques generally belong to one of the following two categories:

- **Simple**—One knows what one is looking for. The first category typically has a well-defined rule or threshold and looks for violations (e.g., all transactions with monetary value larger than a certain threshold or all retired employees who continue to have access to IT systems). The first category of analytics usually employs queries to a database or spreadsheets. Audits use this category of analytics extensively. As data size increases, auditors often rely on aggregated data that IT prepares. Such data may be inadequate for reasons of flexibility and dependence on IT. Data do not need to be big to be useable or useful.

- **Advanced**—One does not know *a priori* what one is looking for (e.g., auditors are not checking whether thresholds are violated or even the threshold values). For example, auditors discover a new phenomenon that is not yet covered by known rules and thresholds. Auditors may be interested in trends or patterns, or they may be interested in discovering new things. The data are often telling a story and, in this category, auditors want to be able to read the story. An example is fraud—auditors may not know exactly if fraud exists and precisely what it consists of because new forms of fraud may appear. Auditors may even be interested in teaching a computer how to read data and make inferences, although the computer's performance should be supervised.

The first category of data analytics is analogous to learning to drive by learning the rules (e.g., how to start the engine, how to brake, how to turn the wheel, understanding speed limits), and the second category is similar to learning to drive by watching videos categorized as good and bad driving. The techniques in the second category are widely used in many fields and are often combined with methods from the first category or other methods from the second category. The main focus of this article is advanced data analytics.

The Complexity of Advanced Data Analytics

Advanced data analytics deals with complex cases that cannot be labeled with a simple rule such as "if the transaction value is larger than a given amount and no prior history of such a transaction by this user is found, classify it as suspicious." These simple rules typically involve thresholds, and crossing these thresholds is an indicator. Sophisticated fraud schemes often evade detection by the simple rules of the first category of data analytics techniques. Advanced data analytics techniques aim to detect these interesting cases. For example, although short duration calls may not be suspicious by themselves, a combination of such calls with other information can be a sign of abuse in telecommunications or private automatic branch exchange (PABX) systems. In general, although an undetected intrusion or fraudulent activity may not violate a single rule or threshold and, thus, evade the first category of analytics, the activity must, nevertheless, exhibit characteristics that are different from normal activity to be detected by advanced data analytics. Advanced data analytics can detect deviations from normal behavior even if normal behavior has not been defined in terms of rules or thresholds. However, to detect these cases, all relevant information, (i.e., fields) must be identified and included in the data, even though it may not be clear yet how the information must be correlated to identify deviations for a fraud case, for example.

The Case for Domain Expertise

Regardless of the data analytics category or method, domain expertise is vital to data analytics and is the prime reason why enterprises recruit new auditors who have domain expertise in a relevant field such as IT or finance.

> **"Regardless of the data analytics category or method, domain expertise is vital to data analytics."**

Domain expertise is required to identify the relevant fields in the data. Systems and data analytics tools return noise if they are provided with irrelevant data, and the cost of investigating false positives is typically substantial. For example, if an enterprise employs data analytics to identify possible fraud, money laundering or a possible attack, a data

scientist can understand data analytics methods and apply them well, but does not necessarily know the relevant fields and how they should be used. A domain expert understands the information that is relevant, or potentially relevant, to fraud, money laundering, an attack, intrusion, etc., but does not necessarily know the data analytics methods for using this information in complex cases.

Does One Need to Be a Data Scientist to Use Data Analytics Tools?

The short answer is no. Ideally, one should be able to instruct a system or tool to, "run method A on data set B, and provide the results." Numerous tools can help auditors do that. The "Top 10 Data Analysis Tools for Business"[5] provides a list of data analytics tools. Most of these tools provide the methods that are described later in this article. The main differences among these tools are ease of use, interfacing and pricing.

Users of data analytics tools must be able to:

- Understand what method A does
- Prepare data set B so that it is useable by method A
- Interpret the results

To be able to use these tools, some familiarization with data analytics jargon and terminology may be required because the methods and submethods often have technical names, such as sequential minimal optimization (SMO), a support vector machines (SVM) method and K-means (the most widely used clustering algorithm).

Data Preparation

Typically, a data set requires data preparation if it contains:

- More than one field (e.g., monetary value and number of transactions)
- A non-numeric categorical field (e.g., male/female)
- A nominal field, e.g., position in the company (administrator, director, data entry personnel)

Data preparation provides the relative importance of each field to programs or tools, e.g., the importance of a common user making 10 transactions vs. an administrator making 10 transactions. Another example is the number of bank transactions made vs. the total amount of the transactions. Are they equally important? Is the total amount more important? If so, how much more important? The data preparation task is akin to defining a common scale to measure different quantities and requires domain expertise. This task may be further complicated if the data set contains non-numeric data, such as yes/no fields that answer questions such as, "Is there a suspicious destination of money transfer?" The non-numeric data must not only be converted to a number, but also to a number that is scaled to assign its relative importance with respect to other fields.

Assigning relative importance numerically is necessary because many methods use the concept of distance, i.e., a measure of how close two events are to each other in their characteristics, e.g., field values for transactions. Each event consists of a number of fields, and each field value must be numeric (or converted to a number) and scaled to reflect its importance with respect to other fields. This is where domain expertise comes into play. No program is smart enough to determine relative importance, unless it is told how to do so.

Data Analytics Methods

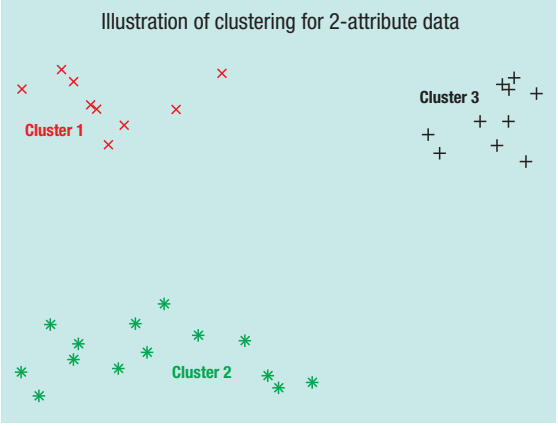
Although more methods are available, there are five data analytics methods that can enhance audits.

Clustering
Clustering organizes data into similar groups, for example:

- A group of managers that shows a similar behavior in outsourcing work that is quite distinct from all other managers
- A group of customers that exhibit a similar behavior, such as high volume transactions of small individual value
- IP packets with special characteristics

Clustering naturally identifies groups with characteristics that are similar within the group and dissimilar from members of other groups. **Figure 1** shows clustering for data with two attributes. Data belong to one of the three clusters shown (X, *, +).

Figure 1—Clustering for Data With Two Attributes



Source: Spiros Alexiou. Reprinted with permission.

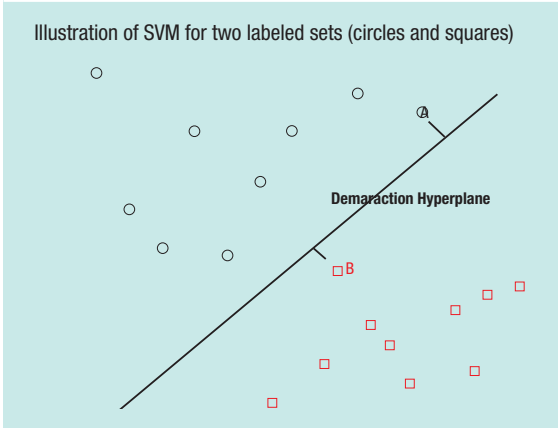
Human analysis and interpretation of the group characteristics, such as center of gravity of the cluster, average values and spread of the data attributes in each cluster, are performed subsequently with the goal of understanding each group. Clustering requires a well-defined distance to access similar behavior. Clustering does not identify strange or suspicious clusters, although it can identify events within a cluster that are distant from most others in the same cluster (outliers). Therefore, humans must interpret and understand the results. Clustering is a very good exploratory tool that makes almost no assumptions and has been used in diverse audits ranging from accounting to network traffic.[6, 7, 8] For example, clustering was applied to network traffic to identify two groups, namely, normal and abnormal network traffic flows.[9] Each member of these groups has characteristics, specifically, packets, bytes and different source-destination pairs, that are closer to the members of the group than to the members of the other group.

Support Vector Machines

The support vector machines (SVM) data analytics method is similar to clustering, because SVM defines, as accurately as possible, the borderline between different clusters, such as fraud/no fraud or solvent/nonsolvent. The feature that separates SVM from clustering is that SVM uses previously labeled data sets to teach the computer to draw the borderline, which, in mathematical terms, is the hyperplane. SVM defines this hyperplane/borderline so that it best divides the two labeled data sets. The division effectively maximizes the area, i.e., the sum of the distances of the closest point of each data set to the borderline, between the two data sets, as illustrated in **figure 2**. Thus a new event, or point, to the left of the established borderline is classified as the rest of the points to the left of the borderline (e.g., fraud/no fraud, positive opinion/negative opinion of a new information system).

Figure 2 shows SVM for two labeled data sets (circles and squares). The demarcation hyperplane best divides the two data sets, i.e. it maximizes the sum of the distances of the closest points A and B from the borderline/hyperplane. SVM is a robust method with a solid mathematical basis and is trainable with relatively few data sets. However, the results are not transparent to users. In addition, the method is quite sensitive to the labeling of borderline cases (points A and B in **figure 2**). An incorrect label in the learning/training data can cause erroneous results. Therefore, the SVM method is best to use when one seeks to determine a borderline and has a high degree of confidence in the labeling of the known cases, especially those that are close to the borderline. Example uses for the SVM method are solvency analysis, intrusion detection and verifying financial statements.[10, 11, 12]

Figure 2—SVM for Two Labeled Data Sets (Circles and Squares)



Source: Spiros Alexiou. Reprinted with permission.

Case-based Reasoning

The case-based reasoning (CBR) method attempts to mimic, on a high level, the reasoning of the human brain. A common problem-solving method that is used by doctors, mechanics and lawyers is to find a similar problem and review how it was handled. CBR uses this same process by saving the solutions to problems in a database. New cases reference the similar cases in the database (**figure 3**).

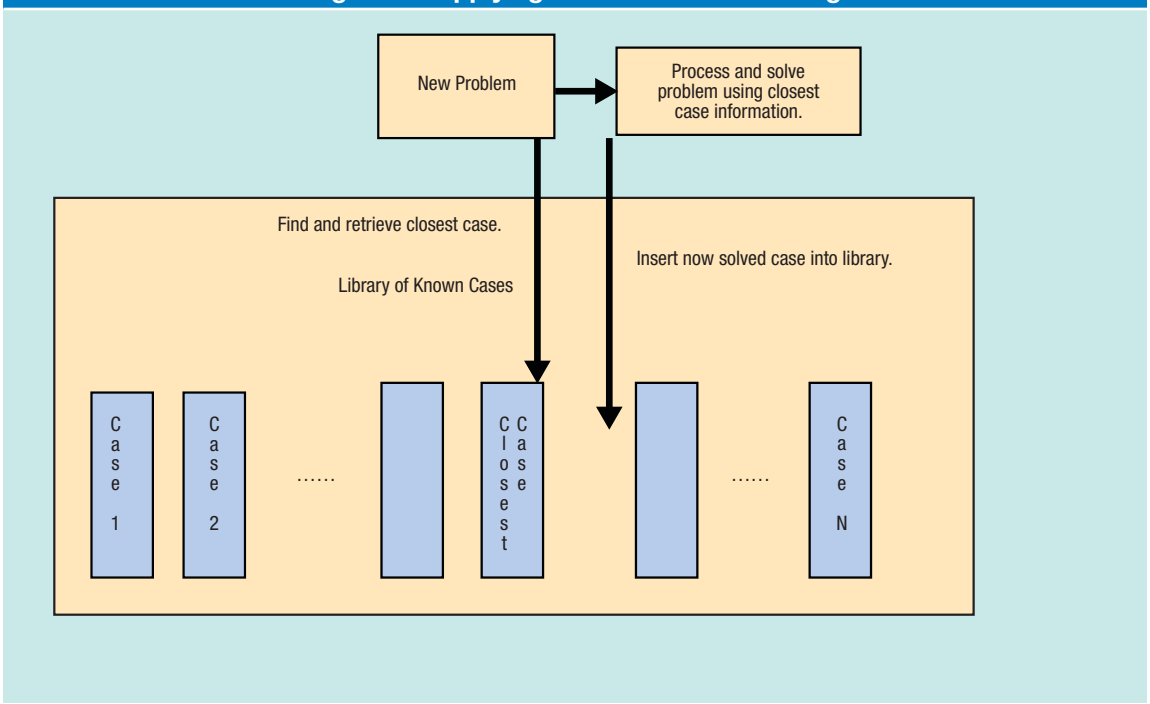
Rules for a new case are constructed based on proximity to the known cases in the database. One weakness of CBR is that a new case that is far from anything known thus far can be misidentified. In practice, the decision or classification is often based not only on the nearest known case, but also on a few nearest neighbors (k-NN), so that the effect of a possible error in a known case is alleviated. The CBR method requires a well-defined distance to access the closeness of two cases. An important advantage of the CBR method is its transparency—the result is based on its similarity to a known case X. Thus, CBR is very useful for classifying a new case based on experience thus far, assuming that previous experience with similar cases exists and their decisions can be explained.

CBR examples in practice range from identifying suspicious transactions to accounting and bank audits.[13, 14, 15, 16, 17] For example, by analyzing the frequency of occurrence of system calls, researchers were able to identify intrusions[18] and, by analyzing access logs, identified anomalous system misuse from inside users.[19]

Artificial Neural Networks

The artificial neural networks (ANN) data analytics method attempts to mimic, on a low-lying neural level, the human brain. Given a set of learning or training data (input), ANN creates a network that produces the known result (output). The ANN method expects that, if the network is given a new set of input data, the network will correctly predict the output. The artificial neural networks method can be viewed as a complex, multidimensional interpolation scheme that, by knowing the output or response to a number of different inputs, predicts the output to different inputs in the same range. The biggest drawback of this method is that it is not transparent to humans and does not provide a simple explanation of why it predicts the output. This drawback is important in many applications, including audits, because it is not acceptable to report

Figure 3—Applying Case-based Reasoning



Source: Spiros Alexiou. Reprinted with permission.

an issue, for example, fraud, which has details that are not understood. Nevertheless, ANN has been used extensively, including for audit purposes.[20] A list of ANN examples in audit, including detection of management fraud using publically available predictors of fraudulent financial statements[21, 22] is available. ANN can be valuable if used as an indicator of something that may be worth investigating.

Random Forest

The random forest data analytics method is a type of decision tree. Decision trees try to create rules from existing evaluated (labeled) cases. For example, one rule that can be deduced is that reporting of financial errors is reduced when an independent audit committee exists and it meets more than twice a year. However, decision trees are prone to overfitting by paying attention to all attributes in the data. For example, a decision tree may use information that is completely irrelevant to the final outcome to formulate a rule. The random forest is an improved variant that uses many different trees that each use a subset of all attributes. The random forest method is designed to alleviate overfitting and the sensitivity of decision trees to noise and uses averaging, which is an effective defense against noise. This method has some similarities to the Delphi method,[23] i.e., an iterative improvement of the opinions of a number of experts that should converge to a single answer. Perhaps a better analogy is a general election or referendum, where most of the voters are assumed to be reasonable on most issues, but each individual voter may have unreasonable views on a few issues. In the same way, the majority of trees in the forest are assumed to be good for most of the data and make different, random errors on some data. If the required answer is a number, then an average of the tree responses is taken as the forest response. If it is a yes/no type of answer, then a majority vote is used. Therefore, a random forest can give humanly understandable rules for classifying current and future cases that are based on already-labeled cases.

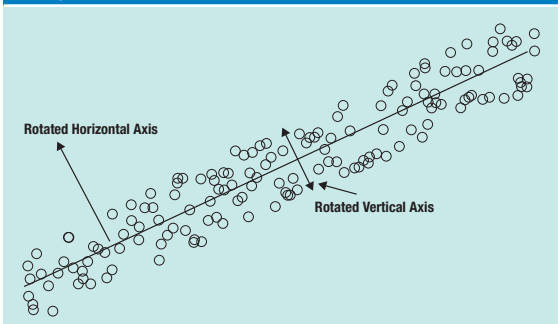
Tools based on random forests typically work off the shelf and give fair results with relatively few data records and many attributes. A recent example of applying random forests to detect financial fraud formed rules based on numerous indicators, such as debt to equity (DEQUTY), current asset ratio (CURAST), and gross profit and EBIT (TPEBIT).[24]

Reduction of Complexity: Principal Axes or Components

Understanding the results in the simplest terms possible is always important, because the results need to be explained to management. Typically, records consist of numerous fields that describe the detailed attributes of an event, e.g., a transaction or login attempt. Principal axes is a mathematical technique to reduce the number of relevant fields. For example, the data analytics methods may detect one type of fraud or another interesting behavior that is characterized by a high number of transactions and low monetary value, and the remaining fields or attributes are largely irrelevant. This example has one principal axis with most of the fraud along this axis. Another axis might describe a different type of fraud and contain a different combination of attributes. This axis is another principal axis.

Figure 4 illustrates the concept of principal component analysis: Data exhibit a much larger variation along the rotated horizontal axis than along the rotated vertical axis. As a result, comparatively little information is lost by ignoring the rotated vertical axis, hence reducing the complexity of the problem to one variable (the rotated horizontal axis) instead of two.

Figure 4—Principal Component Analysis



Source: Spiros Alexiou. Reprinted with permission.

Principal axes analysis aids human understanding, because the large majority of data of interest are along these axes and are easier to understand and visualize. A simple example is intrusions, where the entry and exit time individually might not be relevant, but their difference might be important. Therefore, a different set of axes might be much more informative if it reveals, for example, that intrusions have long durations.

Best of Both Worlds

Methods from both data analytics categories are often combined. Rule-based methods from the first category (one knows what one is looking for) are typically fast, simple and often conclusive. Second category (one does not know exactly what one is looking for) methods are typically more computationally intensive, more complex in data preparation and interpretation, and often indicative. Hence, auditors often apply rule-based methods first and then use second-category methods for cases that are harder to classify.

A significant number of analytics tools are available and many of them are free. These tools can be an important addition to the arsenal of audit tools.

It has been said that, "ANNs and CBR systems have proven they offer better audit effectiveness, better audit quality and reduce audit business risk at a low cost for public accounting firms. It's time these tools are used by auditors."[25] Although every audit is different and has its own requirements, it is likely that many audits could benefit from applying simple and advanced data analytics.

Applying both categories can improve abnormality detection at a low cost, because many of the tools are free and open source. For example, researchers combined their CBR classifier with signature verification to analyze the frequency of occurrence of system calls and identify intrusions.[26] Conventional tools can be used to effectively whitelist cases, therefore, speeding up the procedures. In addition, results from advanced methods can be integrated in rule- and threshold-based methods. For example, traffic flows with certain characteristics corresponding to the abnormal flow cluster will be labeled suspect.

There are data analytics techniques and tools that can significantly aid auditors in discovering knowledge hidden in data, confirming hypotheses and making the most of the available data. These resources are best combined with the auditor's (and possibly other parties') domain expertise and with more conventional tools. The tools are available and many of them free and easy to use once auditors know what they want to do with the data.

"These resources are best combined with the auditor's (and possibly other parties') domain expertise and with more conventional tools."

Endnotes

1 EYGM Limited, "Harnessing the Power of Data: How Internal Audit Can Embed Data Analytics and Drive More Value," EYG no. AU2688, October 2014, *www.ey.com/Publication/vwLUAssets/EY-internal-audit-harnessing-the-power-of-analytics/$FILE/EY-internal-audit-harnessing-the-power-of-analytics.pdf*

2 Izza, M.; "Data Analytics and the Future of the Audit Profession," ICAEW, 22 April 2016, *www.ion.icaew.com/MoorgatePlace/post/Data-analytics-and-the-future-of-the-audit-profession*

3 Rouse, M.; "Data Analytics (DA)," *TechTarget*, January 2008, *http://searchdatamanagement.techtarget.com/definition/data-analytics*

4 *Ibid*.
5 Jones, A.; "Top 10 Data Analysis Tools for Business," KDnuggets, June 2014, *www.kdnuggets.com/2014/06/top-10-data-analysis-tools-business.html*
6 Thiprungsri, S.; M. A. Vasarhelyi; "Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach," *The International Journal of Digital Accounting Research*, vol. 11, 2011, p. 69-84, *www.uhu.es/ijdar/10.4192/1577-8517-v11_4.pdf*
7 Munz, G.; S. Li; G. Carle; "Traffic Anomaly Detection Using K-Means Clustering," 17 January 2016, *https://www.researchgate.net/publication/242158247_Trafc_Anomaly_Detection_Using_K-Means_Clustering*
8 Dhiman, R.; S. Vashisht; K. Sharma; "A Cluster Analysis and Decision Tree Hybrid Approach in Data Mining to Describing Tax Audit," *International Journal of Computers & Technology*, vol. 4, no. 1C, 2013, p. 114-119
9 *Op cit,* Munz
10 Auria, L.; R. A. Moro; "Support Vector Machines (SVM) as a Technique for Solvency Analysis," DIW Berlin, German Institute for Economic Research, August 2008, *www.diw-berlin.de/documents/publikationen/73/88369/dp811.pdf*
11 Abd Manaf, A.; A. Zeki; M. Zamani; S. Chuprat; E. El-Qawasmeh; *Informatics Engineering and Information Science, International Conference*, ICIEIS 2011, Proceedings, Springer, 2011
12 Doumpos, M.; C. Gaganis; F. Pasiouras; "Intelligent Systems in Accounting," *Finance and Management*, vol. 13, 2005, p. 197-215
13 Curet, O.; M. Jackson; "Issues for Auditors Designing Case-based Reasoning Systems," *The International Journal of Digital Accounting Research*, vol. 1, iss. 2, p. 111-123, *www.uhu.es/ijdar/10.4192/1577-8517-v1_6.pdf*
14 Liao, Y.; V. R. Vemuri; "Use of k-Nearest Neighbor Classifier for Intrusion Detection," *Computers and Security*, vol. 21, 2002, p. 439-448
15 Denna, E. L.; J. V. Hansen; R. D. Meservy; L. E. Wood; "Case-based Reasoning and Risk Assessment in Audit Judgment," *Intelligent Systems in Accounting, Finance and Management*, vol. 1, iss. 3, September 1992, p. 163-171
16 Ho Lee, G.; "Rule-based and Case-based Reasoning Approach for Internal Audit of Bank," *Knowledge-Based Systems*, vol. 21, iss. 2, March 2008, p. 140-147, *http://dl.acm.org/citation.cfm?id=1344916*
17 Singh, A.; S. Patel; "Applying Modified K-Nearest Neighbor to Detect Insider Threat in Collaborative Information Systems," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 3, iss. 6, June 2014, p. 14146-14151
18 *Op cit,* Liao
19 *Op cit,* Singh
20 Chao, H.; P. Foote; "Artificial Neural Networks and Case-based Reasoning Systems for Auditing," *Accounting Today*, 2 July 2012, *www.accountingtoday.com/news/artificial-neural-networks-case-based-reasoning-auditing-63178-1.html*
21 Koskivaara, E.; *Artificial Neural Networks in Auditing: State of the Art*, Turku Centre for Computer Science, February 2003, *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.67.459&rep=rep1&type=pdf*
22 Fanning, K. M.; K. O. Cogger; "Neural Network Detection of Management Fraud Using Published Financial Data," *Intelligent Systems in Accounting, Finance and Management*, vol. 7, 1998, p. 21-41
23 Rand Corporation, Delphi Method, Rand.org, *www.rand.org/topics/delphi-method.html*
24 Liu, C.; Y. Chan; A. Kazmi; S. Hasnain; H. Fu; "Financial Fraud Detection Model Based on Random Forest," *International Journal of Economics and Finance*, vol. 7, iss. 7, 25 June 2015, p. 178-188, *https://mpra.ub.uni-muenchen.de/65404/*
25 *Op cit,* Chao
26 *Op cit,* Liao