Ingeniería de Software e Ingeniería del Conocimiento:

Tendencias de Investigación e Innovación Tecnológica en Iberoamérica

Editores:

Raúl A. Aguilar Vera
Julio C. Díaz Mendoza
Gerzon E. Gómez Cruz
Edwin León Bojórquez

ISBN: 978-607-707-096-2

JIISIC'10

Mérida, Yucatán, México

UADY

FACULTAD DE MATEMÁTICAS

"Luz, Ciencia y Verdad"

Ingeniería de Software e Ingeniería del Conocimiento:

Tendencias de Investigación e Innovación Tecnológica en Iberoamérica

Editores:

Raúl A. Aguilar Vera
Julio C. Díaz Mendoza
Gerzon E. Gómez Cruz
Edwin León Bojórquez.

ISBN: 978-607-707-096-2

Alfaomega Grupo Editor

Datos catalograficos

Tendencias de Investigacion e innovacion Tecnologica en Iberoamerica

Alfaomega Grupo Editor, S. A. de C. V., México

ISBN: 978-607-707-096-2

Tendencias de Investigación e Innovación Tecnológica en Iberoarnérica

Derechos Reservados Alfaomega Grupo Editor, S. A. de C. V., México

Alfaomega Grupo Editor, noviembre de 2010

Pitágoras 1139, Col. Del Valle, 03100, México, D. F.

Miembro de la Camara Nacional de la lndustria Editorial Mexicana

Registro No. 2317

Pag. Web: **http://www.alfaomega.com.mx**

E-mail: **atencioncliente@alfaomega.com.mx**

ISBN: 978-607-707-096-2

Nota importante:

La información contenida en esta obra tiene un fin exclusivamente didáctico, y por lo tanto, no está previsto su aprovechamiento a nivel profesional o industrial. Las indicaciones técnicas y programas incluidos, han sido elaborados con gran cuidado por el autor y reproducidos bajo estrictas normas de control. ALFAOMEGA GRUPO EDITOR, S.A., DE C.V. no será jurídicamente responsible por: errores u omisiones, daños y perjuicios que se pudieran atribuir al use de la información comprendida en este libro, ni por la utilización indebida que pudiera dársele.

Impreso en México. Printed in Mexico.

Empresas del grupo:

México: Alfaomega Grupo Editor. S. A. de C. V. - Pitágoras 1139, Col. Del Valle, México, D.F.- C.P. 03100, Tel. (52-55) 5089-7740 - Fax: (52-55)-5575-2420/2490. Sin costo: 01-800-020-4396 E-mail: atencionalcliente@alfaomega.com.mx

Colombia: Alfaomega Colombiana, S. A. - Carrera 15 No. 64 A 29 - PBX (57-1) 2100122, Bogotá, Colombia, Fax: (57-1)-6068648 E-mail: scliente@alfaomega.conn.co

Chile: Alfaomega Grupo Editor, S. A. - General del Canto 370 - Providencia, Santiago, Chile
Tel. (56-2) 235-4248 - Fax: (56-2)-235-5786 - E-mail: agechile@alfaomega.c1

Argentina: Alfaomega Grupo Editor Argentina, S. A. - Paraguay 1307 P.B. "11", Buenos Aires, Argentina, C.P. 1057 - Tel.: (54-11) - 4811-7183/8352, E-mail: ventas@alfaomegaeditorcom.ar

Índice

Pág.

Identification of Noisy Data in Databases by Means of a Clustering Process

H. Kuna, S. Caballero, A. Rambo, E. Meinl, A. Steinhilber, G. Pautch, D. Rodríguez, R. García-Martínez, F. R. Villatoro

Departamento de Informática, Facultad de Ciencias Exactas Químicas y Naturales, Universidad Nacional de Misiones
Grupo de Investigación en Sistemas de Información. Departamento Desarrollo Productivo y Tecnológico. Universidad Nacional de Lanús
Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga

hdkuna@unam.edu.ar

Abstract. Information has become one of the most important assets companies need to protect. From this fact, the audit of systems has a central role in preventing risks related to information technology. In general, development and implementation of the computer-assisted audit technique (CAATs) is still incipient. Data mining applies in an embryonic and asystematicway to tasks related to systems audit. This work tries to show through a case study that the reprocessing stage needed to explore data in an automatic way, specifically the cleaning of data also called data cleaning may be used in the process of auditing systems using clusterization to detect noisy data in databases.

Keywords: systems audit, data mining, software engineering, computer-assisted audit technique, clusterization.

1. Introduction

The effective administration of Information Technology (IT) is a critical element for the survival and success of companies. There are several reasons producing this criticality, for example, dependence of organizationson information for their operation, level of investment in the area of IT and the potential of IT to transform organizations.There is an increasingly stronger relationship between the strategic objectives of a company and IT, a suitable system of internal controls should be implemented to allow the protection of all the elements related to IT, staff, facilities, technology, application systems and data.This makes it increasingly necessary for all organizations and not only the large ones, the

ensure compliance with the rules and procedures for the handling of the policies related to Information Technology.
Systems audit is the systematic process which allows for collection, analysis and evaluation of evidence, to determine in a certain way if the information systems meet effectively and efficiently the objectives of the organization, if the information generated has acceptable levels of quality and safety or if resources are used efficiently. Systems audit has become a central element in the protection of the assets of organizations related to IT. In order to carry out his task the systems auditor uses various techniques to obtain evidence. Given the complexity of information systems, it is necessary for the traditional task performed by the auditor (interviews, revisions, observation, etc.), to be supplemented with the use of specific software which helps to make the process of obtaining evidencemore efficient and objective. The computer-assisted audit technique is the set of programs and data used by auditors in their tasks.
Data mining aims to obtain automatically patterns of behavior in large databases. There is some research related to the use of data mining in the audit systems, even though it is partial and inadequate. Undoubtedly, this type of tool enables the systems auditor to makehis task much more efficient.

The aim of this work is to present the state of the art related to the use of data mining in the process of auditing systems, and through a case study to try to demonstrate the possibility of isolating in clusters noisy data, both discreet and continuous in large databases. Section 2 describes the state of the art in relation to systems audit and data mining. Section 3 describes the materials and methods used. Section 4 contains results and their interpretation. Section 5 contains conclusions and future research, and the last sectionpresents references of the paper.

2. Auditing Systems and Data Mining

In this section we present basics of auditing (section 2.1), data mining (section 2.2) and data mining and auditing systems (section 2.3)

2.1. Basics of Auditing

The audit system is the set of techniques, activities and procedures designed to analyze, evaluate, monitor and recommend on issues related to planning, monitoring, efficacy, safety and adequacy of information systems in the enterprise [1]. Auditors of information systems [2] review and evaluate the development, implementation, maintenance and operation of the components of automated systems and their interfaces with external and non-automated systems.

The audit of internal control systems aims to examine and evaluate the quality and adequacy of controls established by the entity to achieve better performance. The auditor should inform and shape opinion about the reasonableness of such controls, giving an account of items found and recommending proposals for improvement.

SAP Standard 1009 [3], called Computer Assisted Audit Techniques (CAATs) raises the importance of using CAATs in auditing systems. SAP Standard 1009 defines them as computer and data programs used by the auditor as part of the audit procedures to process data of significance in an information system. ISACA [4] has developed the G3 guidelines regarding the use of CAATs.

At national and international level, there are different norms which try to standardize the process of the systems audit. Some of them are explained below.
The mission and objectives of COBIT (Control Objectives for Information and Related Technology) [5] is to research, develop, publish and promote a set of control objectives in information technology (IT) with authority, to date,at international level and generally accepted in daily use by company managers and auditors. The Information Systems Audit and Control Foundation [5] and the sponsors of COBIT have designed this product primarily as a source of instruction forsystems auditors. COBIT has been developed to improve standards and practices of control and IT security to provide a framework for management, users and auditors.

There are some ISO standards [6] related to information security such as ISO 27001 and ISO 17799 to complement the good practice developed in COBIT.

The Information Systems Audit. and Control Association (ISACA) has established a set of general rules for the auditing systems of information.

2.2 Data Mining

Data mining [7] is defined as the process by which understandable and useful knowledge which was previously unknown from databases in various formats, is extracted automatically. Data mining is a key element of a broader process that aims to Knowledge Discovery in Databases (KDD) [8]. This process has a first step data preparation, then the process of data mining, the obtention of patterns of behavior, and the evaluation and interpretation of the patterns discovered.

There are different types of models; these can be of two types, descriptive or predictive.

Predictive: This model aims to estimate unknown values of variables of interest.

- Classification, the goal is to predict which class a new instance of a database belongs to, where the attributes can take discrete values.
- Regression, in this case the predicted value is numeric.

Description: explores the properties of the data examined in order to generate labels or groupings.

- Clustering is to analyze data to generate labels.
- Correlation is used to determine the degree of similarity of the numerical values of two variables.
- Rules of Association aims at finding non-explicit relationships between attributes, it is typically used in the analysis of the contents of a shopping cart.
- Sequential association rules are used to determine sequential patterns in data based on time.

Many authors consider data mining as a step in the process of discovering knowledge, this process involves the following stages [9]: selection of data (data selection), integration of data (data integration), clean data (data cleaning), transformation of data (data transformation), data mining (data mining), assessment of patterns (pattern evaluation) and presentation of knowledge (knowledge presentation).

There are different paradigms behind the techniques applied in the process of data mining: techniques based on decision trees and rule learning systems, algebraic and statistical techniques, techniques based on artificial neural networks, technology-based learning instances or cases, evolutionary algorithms, bayesian techniques. There are other techniques such as stochastic, relational, statements, etc. and a variety of hybrid techniques.

2.3 Data Mining and Auditing Systems

The greatest development of the use of data mining activities related to the audit system is related to intrusion detection networks. Given the objective of this work, the main applications are related to the use of data mining systems applied in systems audits:

- Continuous Auditing and Data Mining
- Research Review in finance and accounting DM
- Analysis of the audit log DM
- Companies in progress and in financial distress
- Assessment of risk control.
- Customer frauds in credit card customers.

Detect noises in databases implies that an audit requires the auditor's need to conduct further tests in order to verify the possible reasons creating such noise and to be able to convert those tracks into evidence. Many are the reasons which can generate this noise [11] [12], for example errors in the analysis of errors in the design of the interface, errors in the design of the database, unauthorized access to the database, etc. In general, previous experiences identified in the use of data mining in Systems Audit, were not related to the detection of such problems, the literature shows experiences primarily with the detection of intrusion in networks and of different types of frauds.

3. Materials and Methods

The objective is to verify that it is possible to identify, through a process of grouping noisy data in databases in order to isolate the cluster having this feature. In accordance with the objectives posed data mining proposes various possible methodologies to apply to the specific problem addressed. Table 1 shows the correlation between the processes, technologies and goals in data mining. For experiments using open source tools, Table 2 shows the products and versions which were used. Table 3 shows overall clustering algorithms used.As a proof of concept, the Mushrooms database in the Machine Learning Repository at UCI [13], was used, Table 4 shows the structure of the database.

Noise was introduced on the following attributes:

- `cap-shape_c` (continuous attribute)
- `stalk-shape` (discrete attribute)

Types of noise:

- Noise type 1: Mean values of the attribute `cap-shape_c` is 8 tuples were introduced with a range of 30 to 100.
- Noise Type 2: 8 tupleswere introduced with a range of 1000 to5000 in the `cap-shape_c.`
- Noise Type 3: 12 tuples were added to noise in the `stalk-shape,` data with noise added have double the number average of characters of that attribute in database.

•

Table 12. Correspondence between processes, technology and objectives in DM

Data Mining Processes	**Best Applicable Technology**	**Goal**
Prediction	Back Propagation network	Prediction of attribute values
Clustering	Self-organizing maps (SOM) / W-CLOPE / Bregman	
Induction	Top Down Induction of Decision Trees (TDIDT)	Discovery of behavior rules
Weighting	Bayesian Nets	Weighting the interdependence of attributes
Clustering + Induction	SOM+TDIDT	Belonging to groups rules Discovery
Induction + Weighting	TDIDT + Bayesian Nets	Weighting of relevant attributes in behavior rules
Clustering + Weighting	SOM + Bayesian Nets	Weighting of relevant attributes in each groupdiscovered

Table 13. Software Used

Product	**Version**
Tanagra	1.4.25
Weka	3-6-0
RapidMiner	4.4.000

Table 314. Algorithms Used

Tool	Technique	Tool	Technique
RapidMiner	Bregman	WEKA	W-FarthestFirst
RapidMiner	W-CLOPE	WEKA	W-SimpleKMeans
WEKA	W-Cobweb	Tanagra	SOM
WEKA	W-EM		

Table 415. Structure of the Database (5417 tuples)

Attribute	type of attribute	Attribute	type of attribute
cap-shape	discrete	stalk-surface-below-ring	Discrete
cap-shape_c	continuous	stalk-color-above-ring	Discrete
cap-surface	discrete	stalk-color-below-ring	Discrete
Bruises	discrete	veil-type	Discrete
Odor	discrete	veil-color	Discrete
gill-attachment	discrete	ring-number	Discrete
gill-spacing	discrete	ring-type	Discrete
gill-size	discrete	spore-print-color	Discrete
gill-color	discrete	population: abundant	Discrete
stalk-shape	discrete	Habitat	Discrete
stalk-root	discrete	Class	Discrete
stalk-surface-above-ring	discrete	---	---

4. Results and Interpretation

Tests were conducted (results are shown in Table 5), where the column "Number of Clusters" represents the total number of clusters generated by the algorithm. For the experiment, this amount was increased, beginning with two clusters increased in two more to achieve the goal and if with 50 clusters the isolation of noise was not achieved, the result of the test was considered as negative.

Table 516. Results Obtained

Algorithm	Total Number of Clusters	Clusters generated with most or all noisy data		
		Noise Type 1	Noise Type 2	Noise Type 3
W-CLOPE	46	X	X	X
Bregman	47	X	X	--
SOM	4	X	X	--
W-Cobweb	50	--	--	--
W-EM	50	--	--	--
W-FarthestFirst	50	--	--	--
W-SimpleKMeans	50	--	--	--

The results of the experiments with the test data used showed that the algorithms SOM, BREGMAN and W-CLOPE had the best results to isolate in a cluster the continuous noise data, clusters were created containing most or all of the noisy data introduced into the database. In the case of discrete noise data showed the best result W-CLOPE.
For other algorithms, continuous and discrete noise data were divided into several clusters, was not isolated all or most of the tuples in a cluster with noise.

- For continuous data , clusters with noisy data were isolated both those closest to the average of the attribute and those most distant.

  In the case of SOM deemed very good results with continuous data a cluster was generated for all noisy data within the range of 30/100 and another cluster with the all noisy data within the range of 1000/5000 of the attribute `cap-shape_c,` both clusters contain only noise data.

  In the case of BREGMAN with continuous data a cluster was generated for all noisy data within the range of 30/100 and another cluster with 6 tuples of noisy data within the range of 1000/5000 of the Attribute `cap-shape_c,` both clusters contain only noise data. The other 2 tuples the range of 1000/5000, were included in another cluster that contains the majority of data without noise.

  In the case of W-CLOPE with continuous data a cluster was generated with 5 tuples for noisy data within the range of 30/100 and 3 tuples with the noisy data within the range of 1000/5000 of the attribute `cap-shape_c.` The other 3 tuples with the noisy data within the range of 30/100and the 5 tuples noisy data within the range of 1000/5000, were included in other clusters that contains the majority of data without noise.

- In the case of the attribute `stalk-shape`(discrete data) acluster was generated with algorithm W-CLOPE containing 9 tuples with noisy data, the other 3 tuples with noisy data were included in other clusters that contains the majority of data without noise.

5. Conclusions and Future Research Lines

CAATs tools are essential for auditing systems, they automate and objectify the collection of audit trails, which as from this identification make it possible the obtention of evidence to substantiate the findings.Data mining is very useful in audit processeswhen finding patterns of not expectedbehavior in DBs.The experiment demonstrated that it is possible to isolate clusters in noisy data, including continuous and discrete data, while it is necessary to continue experimenting with different types of noise, in particular the best result was obtained with the continuous noise data applying SOM, thus providing an important clue to systems auditors, given that carrying out this work "manually" would imply a time and expertise which are not always available.

The detection of inconsistent data by means ofthe use of information and the development of specific procedures for the application of data mining to systems audit appear as the main research lines to follow

6. Referencias

1. Rivas J.: Auditoría Informática, Ediciones Díaz de Santos, ISBN 84-87189-13 (1998)
2. Information System Audit and Control Association. Available at:http://www.isaca.org
3. Statement of Auditing Practice. Available at:http://www.eisi.ues.edu.sv/jportillo/taac.doc
4. Information Systems Audit and Control Association. Available at:http://www.isaca.org
5. Control Objectives for Information and related Technology. Available at:http://www.isaca.org/cobit/
6. Asociación de Auditoría y Control de Sistemas de Información. Available at:http://www.adacsi.org.ar
7. Clark P., Boswell R.: Practical Machine Learning Tools and Techniques with Java Implementation. Morgan Kaufmann Publishers (2000)
8. Fayyad U.M., Piatetsky-Shapiro G, Smyth P.: From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data Mining, pp:1-34, AAAI/MIT Press (1996)
9. Britos P., Hossian A., Garcia Martinez R., Sierra E. Mineria de Datos basada en Sistemas Inteligentes, Editorial Nueva Librería. ISBN 987-1104- 30-8 (2005)
10. Garrity Edward J., O'Donnell Joseph B., Sanders GL.: Continuos Auditing and Data Mining. Encyclopedia of Data Warehousing and Mining, pp 217-222, Idea Group Reference (2006)
11. Kuna, H., García Martínez, R., Villatoro, F. R. Pattern Discovery in University Students Desertion Based on Data Mining. Book of Abstracts of IV International Meeting on Dynamics of Social and Economic Systems. Pág. 48 (2009)

12. Kuna, H., García-Martínez, R., Villatoro, F. R.Procedimientos de la Explotación de Información para la Identificación de Datos Faltantes, con Ruido e Inconsistentes. Proceedings XI VICC-2009: 236-238. ISBN 978-950-605-570-7 (2009)
13. UCI. Machine Learning Repository.Available at:http://archive.ics.uci.edu/ml/datasets/Mushroom

Alfaomega Grupo Editor