Risk based Government Audit Planning using Naïve Bayes Classifiers

Remis Balaniuk[1,2], Pierre Bessiere[3,4], Emmanuel Mazer[4], Paulo Cobbe[1,5]

[1] MGCTI,,- Catholic University of Brasilia, SGAN 916 – Módulo B – Asa Norte, cep:70790-160, Brasília DF

[2] Tribunal de Contas da União, Setor de Administração Federal Sul, SAFS quadra 4, lote 1, cep:70042-900 Brasília – DF, Brazil

[3] LPPA - Collège de France, 11 place Marcelin Berthelot, 75231 Paris cedex05

[4] CNRS, E-Motion, LIG - INRIA, 655 avenue de lÈurope, 38334 Montbonnot, France

[5] Information Technology Department, UniCEUB College, SEPN 707/907 Asa Norte, cep:70790-075 Brasília – DF, Brazil

remis@ucb.br, pierre.bessiere@college-de-france.fr, emmanuel.mazer@inria.fr, paulocobbe@yahoo.com

Abstract. In this paper we consider the application of a naïve Bayes model for the evaluation of fraud risk connected with government agencies. This model applies probabilistic classifiers to support a generic risk assessment model, allowing for more efficient and effective use of resources for fraud detection in government transactions, and assisting audit agencies in transitioning from reactive to proactive fraud detection model.

Keywords: Data mining, naïve Bayes, fraud detection.

1 Introduction

Computer technology gives auditors a large set of techniques for examining the automated business environment. Computer-assisted auditing techniques (CAATs) have become standard practice in corporate internal auditing.

Audit software permits auditors to obtain a quick overview of the business operations and drill down into the details of specific areas of interest.

Audit software can also highlight individual transactions that contain characteristics often associated with fraudulent activity. A 100% verification can identify suspect situations such as the existence of duplicate transactions, missing transactions, and anomalies.

Auditors usually look for patterns that indicate fraudulent activity. Data patterns such as negative entries in inventory received fields, voided transactions followed by

"No Sale," or a high percentage of returned items may indicate fraudulent activity in a corporation. Auditors can use these data patterns to develop a "fraud profile" early in their review of operations. The patterns can function as auditor-specified criteria and transactions fitting the fraud profile can trigger auditor reviews [1].

Nevertheless, when it comes to government auditing, the scale and the complexity of the roles to be considered can prevent the use of most methods and technologies so successful on the corporate world.

As stated by the US Comptroller General[2], the major roles and responsibilities of governmental auditing are: combating corruptions, assuring accountability, enhancing economy, efficiency, and effectiveness, increasing insight, facilitating foresight.

Corruption can exist within government or on the part of contractors and others who conduct business with government. Fighting corruption requires a strong and effective audit function. For that to occur, government audit agencies must be assured free access to process, routines and records of government organizations.

Government auditing can also add value by analyzing the efficiency and effectiveness of government organizations, programs and resources.

Compared to corporate internal auditing, government auditing acts on a much larger universe, with concerns ranging from auditing specific operations or contracts to evaluating program effectiveness and performance.

As indicated by a survey by the Global Audit Information Network [3], the top challenge facing all government audit organization is adequate audit staffing. The large audit universe, the diversity and complexity of the topics being covered makes it impossible for audit agencies to perform a 100% verification on all government operations or entities. The same survey also indicated the ability to plan based on risk as being part of the top ten challenges imposed to all government audit organizations.

Traditional methods for audit planning are usually based on management requests, auditors' experience or expertise or simply on statutes or regulations. These methods can work fine inside a corporation but are much less effective at the government level.

Risk-based audit planning, on the other hand, can result in disciplined analytical approaches to evaluate the audit universe, highlights potential risks that might otherwise be unknown, fosters dedicated audit coverage to high-risk areas, and allocates resources where pay-back is greatest [4].

A risk assessment process for audit planning proposed by the Institute of Internal Auditors (IIA) [4] is based on five steps:

- Define the audit universe
- Identify and weight risk factors
- Establish a mechanism and score risk factors for auditable units
- Sort the auditable units by total risk score
- Develop the audit plan based on the ranked audit universe.

The aim of this paper is to propose a method, based on a probabilistic classifier, to support this generic risk assessment process. Our method can score auditable units using a formally defined mathematical framework. Extensive databases containing records from government operations can be combined to auditors knowledge, fraud profiles, impact factors or any other relevant metric in order to rank an audit universe.

2 Related work

2.1 Fraud detection and Risk analysis using data mining

Data mining has become an increasingly popular tool used by business, organizations and governments for aiding audit professionals in risk analysis and fraud detection. Several scholarly works have been written on application techniques, particularly for such businesses as financial and security exchange institutions, telecommunications and insurance companies, whom, along with their clients, incur incalculable financial losses due to fraud every year around the world [5][6].

When properly applied, data mining techniques are able to identify trends that indicate suspicious or fraudulent activities, casting light on transactions hidden among the crowd. By reducing the universe of transactions or activities to a smaller subset, data mining allows decision makers to concentrate their efforts on higher risk transactions, and act to mitigate the repercussion of fraudulent activity in affected organizations [5][7][8].

Risk prediction is an important contribution of data mining to a decision making process. By quantifying the possibility that a given event may occur in the future, it provides the decision maker with a basis for comparing alternative courses of action under uncertainty [9]. Despite of the fact that it is impossible to ascertain with complete certainty events which have yet to take place, risk analysis allows decision makers to define possible outcomes and assess the risks associated with each, and make a decision based on these possible alternatives. This assessment does not shield the decision maker from negative outcomes, but should ensure that positive outcomes are reached more often than not [10].

2.2 Naïve Bayes classifiers and fraud detection

Naïve Bayes is a mining technique not commonly associated with fraud detection in the scientific literature. In their review of the academic literature on data mining applications in financial fraud detection, Ngai et al [5] indicate very few studies that applied naïve Bayes algorithms for this purpose.

That trend contrasts with Viaene, Dering and Dedene [11] findings, who present a successful application of naïve Bayes algorithm to personal injury protection (PIP) claims for the State of Massachusetts. Their findings suggest that this algorithm can lead to efficient fraud detection systems for insurance claim evaluation support.

Viaene et al [12] also indicates that naïve Bayes algorithms showed comparative predictive performance to more complex and computationally demanding algorithms, such as Bayesian Learning Multilayer Perceptron, least-squares support vector machine and tree-augmented naïve Bayes classification, in a benchmark study of algorithm performance for insurance fraud detection.

The emphasis of research of complex unsupervised algorithms presents to Phua et al [7] a problem for the future. The authors suggest that in order for fraud detection to be successfully implemented in real time applications, less complex algorithms, such as naïve Bayes, have to be considered as the only viable options.

3 Description of the method

The method proposed in this paper performs risk evaluation based on classification rules. Rules are built by exploring large databases collected in daily activity of government agencies. Typically, the auditable universe is large and we want to classify auditable units to one of two groups: high risk units and low risk units. Moreover, these units should be sorted with respect to their risk.

3.1 Naïve Bayes classifiers

The probability model for a classifier is a conditional model : $P(C|H_1,...,H_n)$ over a dependent class variable C with a small number of outcomes or classes, conditional on several feature variables H_1 through H_n.

Using the well-known Bayes' theorem we write:

$$P(C|H_1,...,H_n) = \frac{P(H_1,...,H_n|C)P(C)}{P(H_1,...,H_n)}$$

Because the denominator does not depend on C we are usually interested only in the numerator of the right side fraction. The values of the features are also given and consequently the denominator is constant.

The numerator is equivalent to the joint probability model: $P(C,H_1,...,H_n)$.

The problem is that if the number of features is large or when a feature can take on a large number of values, the computation of such a model can be infeasible.

The "naive" conditional independence assumption assumes that each feature is conditionally independent of every other feature: $P(H_i|C,H_j) = P(H_i|C)$.

This strong assumption can be unrealistic in most cases, but empirical studies related to fraud detection show that most frequently the method presents good performance [11]. Zhang [15] explained the superb classification performance of naïve Bayes in real applications, where the conditional independence is rarely true, showing that dependence among attributes usually cancel out each other.

Under these independence assumptions the conditional distribution over the class variable can be expressed as:

$$P(C|H_1,...,H_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(H_i \mid C)$$

where Z is a constant if the values of the feature variables are known. $P(C)$ is called the class prior and $P(H_i|C)$ are the independent probability distributions

By using this naïve Bayesian algorithm it is then possible to obtain a probability distribution of objects belonging into classes. Threshold rules can be used to decide when a probability is strong enough to assign an object into a group. Depending on these rules it happens that an object is not assigned to any group or to more than one group, like in fuzzy logic. Naïve Bayes also naturally deals with missing values, what is difficult to achieve using other methods like decision trees or neural networks. Resulting models are self explainable, unlike other methods like neural networks.

Further information about Bayesian Classifiers can be obtained at [7].

3.2 Adapted Naïve Bayes Classifier

A major difficulty imposed on risk evaluation by government auditing agencies is the lack of consistent fraud databases. Most methods used in AI are based on supervised learning, where a set of examples are used to define an inference model. Neural networks and decision-trees are examples of supervised learning methods.

Detected fraud cases are usually reported in unstructured documents. These are typically disconnected, identified and described in very specific contexts and are not statistically relevant considering the number of "variables" and "states" in which a systemic fraud could be described.

This lack of information compromises the typical use of Bayesian classifiers. Conditional probabilities for the output classes cannot be established without a consistent base of tagged examples distributed on the input space.

In order to overcome this handicap we adapted the standard naïve Bayes classification approach making two assumptions:

1. Statistically, the high risk group is much smaller than the low risk group: fraud is an exception (typically less than 1% of the auditing units)
2. High risk units can be described by rules: auditing experts are able to describe fraud profiles from their field expertise

From these assumptions and having a non labeled extensive database describing our auditing universe, our problem is decomposed in two different steps:

- Find the conditional probabilities for the low risk group: from the first assumption we can consider that the fraud cases inside our large database are not statistically relevant, so we will estimate this conditional probability using all the data contained in the database as if there were no fraudulent units there: $P(H_i \mid C=low\ risk) = P(H_i)$
- Find the conditional probabilities for the high risk group: the audit expert directly defines the shape of a probability distribution based on his expertise.

3.3 The risk assessment process

Distinct risk assessment processes are defined for distinct sets of auditable entities.

The first step in implementing a successful risk based audit process is to clearly define and understand a chosen audit universe.

The adoption of CAATs by government audit agencies enables the compilation of timely, reliable, and meaningful data that can be used for planned audits on an ongoing basis. Based on compiled data, once CAATs are implemented, auditors or auditing systems can routinely review significant financial and non-financial processes and identify potential audit issues.

Once an audit issue is identified the corresponding audit universe will be defined as the set of entities to be classified in order to reflect the risk associated to that issue.

An audit issue can be related to a business process as government purchasing / procurement, general ledger, treasury, payroll, accounts payable, inventory and fixed assets. Broader issues can be related to government programs. More specific issues can be related to public employees or public contractors.

Correspondingly, the audit universe can be a set o purchases, contracts, assets, programs, public employees or contractors.

Once the audit universe is defined the second step is to identify its risk factors. The risk factors of an audit universe are related to the audit issue and the business rules associated to the entities set.

The risk factors will define the conditional features of the probabilistic classifier. Because of the naïve Bayes assumption of variables independence, a careful choice of risk factors must be made considering their independence. As explained by Zhang [15], violations of the independence usually do not compromise the optimality of the model once local dependencies usually cancel out each other. However, if the model does not perform well is still possible to identify which factors are affecting the model performance using analysis of covariance as proposed by Zhang [15].

There is no need to define weight factors while using probabilistic classifiers once the nature of the feature variables define a common framework to all measures.

To illustrate how the choice of the risk factors occurs, consider as audit issue the fraud detection on public purchasing/procurement. Public procurement comprises government purchasing of goods and services required for State activities, the basic purpose of which is to secure best value for public expenditures. In both developed and developing economics, however, the efficient functioning of public procurement may be distorted by the problems of collusion or corruption or both [13].

Considering fraud on public procurement our audit issue, our audit universe will be the set of public purchases, preferably over a large period of time (five to ten years) and a large scope of purchasing organizations, contractors and purchased goods and services. The risk assessment would be ideally performed based on a detailed database describing all relevant aspects of each purchase.

One obvious first risk factor associated to public purchases is the space left for competition between potential sellers. The corresponding high risk rule would be: *less competition = higher risk*. To compute the competition level within a procurement process one could use a heuristic model based on its characteristics. The existence or not of an open call for bids or tenders, the type of procurement and the number of registered bidders can be the input for this model. The heuristics can be based on business rules: restricted or invited tenders are more suitable for fraud; electronic procurement auctions (e-procurement, e-reverse auctioning, e-tendering), on the other hand, tend to be more competitive and less suitable for fraud; the competition of a bid is proportional to the number of registered bidders.

Other risk factors to consider could be:

- Value: more money = higher risk
- Bid protests: more protests = higher risk
- Bid winner profile: no qualified winner = higher risk
- Bid winner previous sanctions: sanctioned winner = higher risk

Some risk factors can require the use of external data, not directly related to the audit universe. The winner previous sanctions, for instance, would require a sanctions database. This a common issue when adopting CAATs. Analytical auditing databases require information covering a broad scope of related themes.

Factors associated with different concerns than the risk itself can be added.

- Political and economic impact: expenditure contributes to the achievement of policy goals = higher economic impact

- Social impact: public health spend = higher social impact
- Timing and effectiveness: older purchases = less auditing effectiveness

The only requirements to include a new factor to the model are the possibility to estimate a corresponding numerical value for each entity belonging to the audit universe and the existence of a rule for the high risk/impact/effectiveness group.

Once the risk factors are identified and computed for the whole audit universe, the next step will be to run the naïve Bayes algorithm and compute the probability distribution of entities belonging into classes.

Because we have only two classes (high and low risk) we obtain two values for each entity: $P(C=high\ risk|H1,...,Hn)$ and $P(C=low\ risk|H1,...,Hn)$.

If a choice of subsets of risky and non-risky entities is required a threshold rule must be defined in order to decide when a probability is strong enough to assign an entity into a class.

Moreover, the probability $P(C=high\ risk|H_1,...,H_n)$ can be directly used to sort the auditable units by total risk score. The ranked audit universe can then be used to develop an audit plan.

As in any knowledge discovery process, the auditing results should be feedbacked to refine the assumptions upon which the whole model was constructed.

4 Experiments

This approach was tested by the *Tribunal de Contas da União* – TCU, the Brazilian Court of Audit. The TCU have been using CAATs in the last five years intensively, and has gathered extensive information about the Brazilian public sector. It receives on an ongoing basis data from IT systems of major government agencies. All relevant data is assembled into a large data warehouse.

This test was done as part of a research project involving the TCU, the *Institut pour la Recherche en Informatique et Automatique*- INRIA and the company Probayes[1].

ProBayes developed the ProBT engine [14], a powerful tool that facilitates the creation of Bayesian models.

A number of risk assessment models were built in order to analyze major audit issues like high risk private contractors, collusion between private contractors and corruption between public bodies and private contractors. The audit issues, the audit universes and the risk factors were designed by TCU auditors.

In the following we detail one of these models.

4.1. Corruption in public procurements

As pointed out by the Organization for Economic Co-operation and Development - OECD Global Forum on Competition Debate on Collusion and Corruption in Public Procurement in October 2010 [13]:

[1] http://www.probayes.com

"Corruption occurs where public officials use public powers for personal gain, for example, by accepting a bribe in exchange for granting a tender. While usually occurring during the procurement process, instances of post-award corruption also arise. Corruption constitutes a vertical relationship between the public official concerned, acting as buyer in the transaction, and one or more bidders, acting as sellers in this instance."

We used the proposed naïve Bayes approach to assess risk of corruption between a public body and private companies awarded with its public contracts.

The audit universe was the set of pairs of public and private parties of all public contracts signed by the Brazilian federal administration between 1997 and 2011, totalizing 795954 pairs.

The risk factors chosen by the auditors were:

- Competition: purchases awarded to restricted or invited tenders without a open call bid were considered risky
- Post-award renegotiations of values rising the initial awarded purchase value raise corruption risk
- Post-award renegotiations of values reducing the initial awarded purchase value reduce corruption risk
- The existence of links between the parties was considered risky: a possible link would be public employees from the public body or their relatives which are or were partners or employees of the private company.

The risk factors were then transformed in numerical features. To reduce the computation effort the features were discretized and their values bounded.

The following features were computed for each pair of *(public body, private company)* from the audit universe:

- `NISLICIT`: number of purchases awarded without an open call bid
  - The final feature value was the natural logarithm of the count of purchase awards (values between 0 and 9)
- `TISLICIT`: total amount of the purchases awarded without an open call bid
  - The final feature value was the logarithm base 10 of the total amount of purchases (values between 0 and 5)
- `NREFOR`: the number of post-award purchase value increases
  - The final feature value was the natural logarithm of the count of purchase value increases (values between 0 and 7)
- `TREFOR`: total amount of the purchases value increases
  - The final feature value was the logarithm base 10 of the total amount of purchase value increases (values between 0 and 6)
- `NANULA`: : the number of post-award purchase value reductions
  - The final feature value was the natural logarithm of the count of purchase value reductions (values between 0 and 7)
- `TANULA`: total amount of the purchases value reductions
  - The final feature value was the logarithm base 10 of the total amount of purchase value reductions (values between 0 and 6)
- `TSESO`: indicates the existence or not of links between the public body and the private company
  - The final feature value was 1 for linked pairs and 0 otherwise.

4.2 Results

The conditional probabilities for the seven features are illustrated on the following Figures. Conditional probabilities for the low risk class, obtained from the purchase database are displayed on the left side of the Figures. Conditional probabilities for the high risk class directly assigned by the audit experts, are displayed on the right.

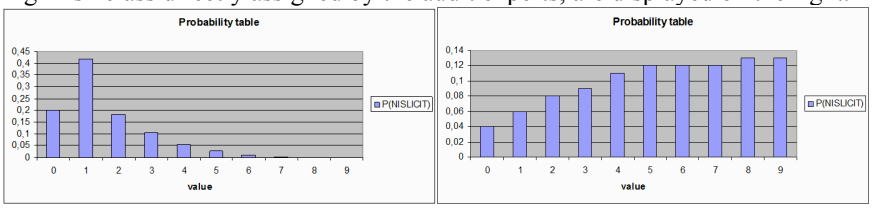

Figure 1: left: *P(NISLICIT□C=low risk)* and right: *P(NISLICIT□C=high risk)*

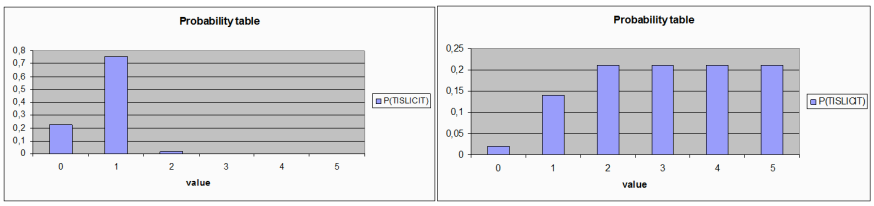

Figure 3: left: *P(TISLICIT□C=low risk)* and right: *P(TISLICIT□C=high risk)*

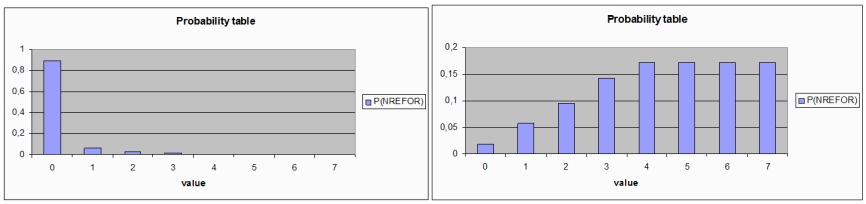

Figure 4: left: *P(NREFOR□C=low risk)* and right: *P(NREFOR□C=high risk)*

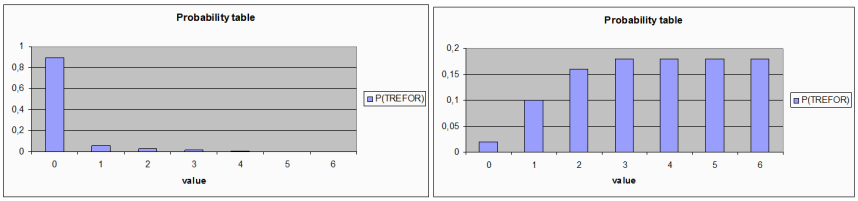

Figure 5: left: P*(TREFOR□C=low risk)* and right: *P(TREFOR□C=high risk)*

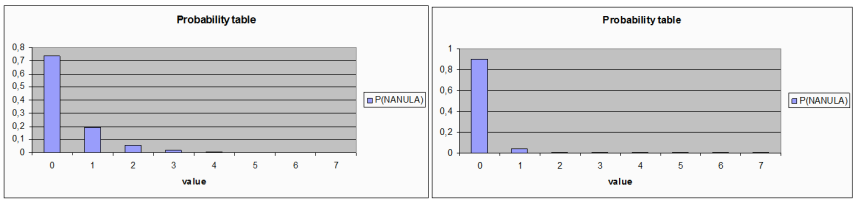

Figure 6: left: *P(NANULA□C=low risk)* and right: *P(NANULA□C=high risk)*

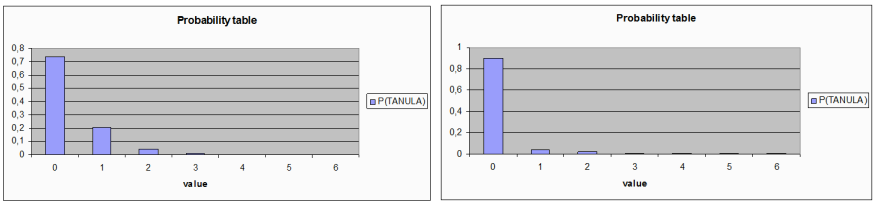

Figure 7: left: *P(TANULA|C=low risk)* and right: *P(TANULA|C=high risk)*

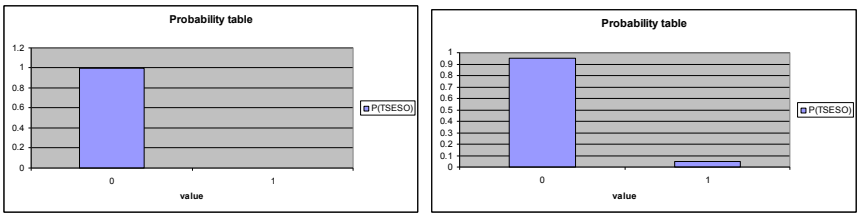

Figure 8: left: *P(TSESO|C=low risk)* and right: *P(TSESO|C=high risk)*

The class prior was assigned an empirical value *P(C) =0.01* based on the auditors expectation of corruption inside the audit set.

The computation of *P(C=high risk|NISLICIT, TISLICIT, NREFOR,TREFOR, NANULA, TANULA, TSESO)* for the set of 795,954 pairs of public and private parties indicate 2,560 pairs with probability higher than 99% (0.99).

This result included parties with obvious links, like private not-for-profit foundations for which the Brazilian public procurement law gives special treatment. Moreover, a number of high risk pairs were known by the auditors from previous investigations where corruption was effectively found. The overall feeling from all auditors to which the high risk list was presented was that the result was very reasonable. Future audit plans will possibly be based on our risk rank.

5 Conclusion

With a growing number of governmental agencies transitioning into unified and consolidated data information platforms, access to information and uniformity are increasing. To deal with all this information government audit organizations are increasingly adopting CAATs in order to routinely review significant financial and non-financial processes and identify potential audit issues.

The steps taken to accumulate and identify significant data elements allow audit organizations to use that data during planned and unplanned audits. The perpetual development of data repositories and implementation of ongoing monitoring can also contribute to the development of periodic government-wide risk assessment.

In this paper we proposed a risk assessment method, based on naïve Bayes classifiers, that can be used by government audit organizations. The proposed method is suitable to the typical risk assessment process for audit planning, as formalized by the Institute of Internal Auditors (IIA).

The main advantages of the proposed method are:

- Integration of auditors knowledge to large data repositories in order to analyze audit issues
- Integration of quantitative risk factors to qualitative aspects to compose probabilistic features
- Natural framework to deal with missing data, data in different scales and from different sources
- Low computational complexity

Upon implementation, this semi-automated risk assessment procedure can help audit organizations transition from a reactive response to a proactive approach to identify and correct issues that may be indicative of fraud, waste or abuse.

References

1. Coderre, D.: Auditing - Computer-Assisted Techniques for Fraud Detection. The CPA Journal, August, http://www.luca.com/cpajournal/1999/0899/0899toc.htm (1999)
2. United States Government Accountability Office – GAO: Government Auditing Standards (The Yellow Book), August (2011), http://www.gao.gov/yellowbook
3. Kelli W. Vito, SPHR, CCP, Auditing Employee Hiring and Staffing, ISBN : 978-0-89413-703-7, Publisher : The IIA Research Foundation, Publish Date : June 2011.
4. The Institute of Internal Auditors: International Professional Practices Framework (IPPF), The IIA Research Foundation (2009)
5. Ngai, E.W.T., Yong Hu, Y.H. Wong, Y.C., Sun X.: The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, pp. 559 – 569. (2011).
6. Panigrahi, S., Kundu, A., Sural, S., and Majumdar, A. K., Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning. Information Fusion 10(4), 2009, 354-363
7. Phua C., Lee V., Smith K., Gayler R.: A comprehensive survey of data mining-based fraud detection research, Artificial Intelligence Review, pp. 1 – 14. (2005)
8. Hormazi, A. M., Giles S.: Data Mining: A Competitive Weapon for Banking and Retail Industries. Information Systems Management; Spring, vol. 21, n.2, pp. 62 – 71. (2004)
9. Nilsen, T., Aven, T.: Models and model uncertainty in the context of risk analysis. Reliability Engineering and System Safety, 79, 309–31. (2003)
10. Nilsen, T.: Foundations of Risk Analysis: A Knowledge and Decision-Oriented Perspective. John Wiley & Sons Ltd, West Sussex, 2003.
11. Viaene, S., Derrig, R. A., Dedene, G.: A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis. IEEE Transactions on Knowledge and Data Engineering, vol. 16, n. 5, pp. 612–620, May. (2004)
12. Viaene, S.; Derrig, R.A.; Baesens, B.; Dedene, G. A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection. J. Risk and Insurance, vol. 69, no. 3, pp. 373-421. (2002)
13. Organisation for Economic Co-operation and Development: Roundtable on Collusion and Corruption in Public Procurement. October, www.oecd.org/dataoecd/35/19/46235884.pdf. (2010)
14. Mekhnacha, K., Ahuactzin, J.M., Bessière, P., Mazer, E., Smail , L.: Exact and approximate inference in ProBT, Revue d'Intelligence Artificielle, Vol. 21/3, pp. 295-332. (2007)
15. Zhang, H.,The Optimality of Naive Bayes, 2004, American Association for Artificial Intelligence (www.aaai.org).