ISSN 2278-8808 SJIF 2014 = 4.889

An International Peer Reviewed & Referred

SCHOLARLY RESEARCH JOURNAL FOR INTERDISCIPLINARY STUDIES

FRAUD DETECTION WITH SOCIAL NETWORK ANALYSIS: ANALYSIS ON IDENTITY FRAUD AND DETECTION SCHEME

Mr. Kishor Keshaorao Wikhe & Arjun Pralhad Ghatule, Ph. D.

Abstract

Fraud is a crime where the purpose is to appropriate money by an illegal form or method. Fraud leads to large losses to businesses or States or individuals. Since its detection is complex, there is not yet a fraud detection framework that can detect and prevent fraud in an efficient& accurate way. Almost any business enterprise that involves money and services can be compromised by fraudulent acts. Such areas as Capital Markets, Social Security, Insurance, Telecommunications, Financial Institutions, and Credit Cards are examples of where fraud may occur and where, in the recent past, there has been an effort to develop methods to combat this type of financial fraud. Each one of the areas has specific characteristics; therefore, a single solution that may be deployed to fight fraud in Credit Cards cannot be applied to Insurance industry. As a consequence, one of the approaches to fight fraud is to pursue a model that describes fraudulent behaviors, or, better, create mechanisms that distinguish fraudulent from non-fraudulent behaviors.

Keywords: *Fraud detection, social network, Organization*

1. INTRODUCTION:

The approaches to find fraud is a model that describes fraudulent behaviors, or, better, create mechanisms that distinguish fraudulent from non-fraudulent behaviors technically, these mechanisms can be created using a data mining classification, making use of a set of historical records, i.e., records of past customers already known as fraudulent and non-fraudulent (the training set). However, applying classification techniques for fighting fraud always deal with a particularity: the existence of an unbalanced training dataset – a set that has many non-fraudulent records and only a few fraudulent cases. Because of this characteristic, the result of applying traditional classification techniques, like decision trees or neural networks, are not enough for obtaining a good classifier.

In the context of classification, there are three major approaches that may be used to create a classifier for fraud detection: balance the data, consider different errors costs and detect outliers. This study describes defines fraud and what the problems associated with classification algorithms, when one attempts to detect and predict fraud. Along with this description, it presents a set of most used techniques to face the stated problems. Lastly, it comments on the particularities of some business areas that are affected by fraud.

The specific problem that will be addressed is this study is fraud detection at the ―DirectorGeral de Impostos(that is in a free translation to the State Taxes Organization) on the Value Added Tax (VAT), in particular on the detection concerning Carrousel fraud cases. Thus, one of the objectives of this study is to develop a fraud classifier, with an acceptable accuracy level for fraudulent cases. Moreover, this must be accomplished by solving or minimizing financial fraud detection issues. In order to obtain this goal, several classifiers will be created, implementing the most relevant approaches for fraud detection, as techniques to balance the dataset or algorithms that learn with different errors costs.

Along with this, it is important to remember that fraud is perpetuated by people or by organizations that cannot detach themselves from the rest of the world. As a result, fraudulent persons and organizations are connected among themselves and to the rest of all other honest and fair organizations. It is, in fact, is important not to deal with entities separately, when looking for fraud but, also, to deal with the relationships between these entities. As an example, the owner of a fraudulent organization is probably the person responsible for the fraud itself. If he owns another organization, it is more likely that this second organization is also fraudulent. Another issue is the fact that some take advantage of carrousel fraud, which can only be accomplished with a link of several organizations. Because of this, the analysis ofsocial networks within fraudulent organizations and its people can be extremely important when searching for fraud.

In this manner, with the work presented here, the social networks of each organization will be analyzed in order to detect patterns that are common to fraudulent organizations. These patterns will enforce the data in VAT declarations, in order to create a dataset with more useful information. The dataset is the base support for the creation of new classifiers, which will be much more accurate than the original ones. With this new dataset that, although having more information about each organization, is still an unbalanced dataset, to which it will be also applied certain techniques do deal with that specific problem.

2. REVIEW OF LITERATURE:

Without a doubt the issues depicted are simply advisers for the distinctive mining approaches that by and by exist to manage the issue of misrepresentation recognition. Specifically, the issues of skewed datasets with minority classes can be taken care of by applying two distinctive methodologies: the calculation or information approaches (Kumar, 2007). The information approach is utilized at the pre-handling period of the KDD procedure. This assignment comprises of re-inspecting the lopsided classes. Re-inspecting information is a method that permits one to make more record cases of a class (over–sampling) or kill a few records of a class (under-testing). There are three structures for accomplishing a more adjusted dataset. That is under-examining the dominant part class, over-testing the minority class, or doing both. The immense preferred standpoint of this strategy is that it can be utilized with any characterization system (Ahumada, 2008). The calculation approach is utilized as a part of the preparing period of the KDD and comprises of changing existing information mining calculations by conforming the expenses of the blunders (Han, 2005). In this way, by doling out a better cost than the false negatives (fake records marked non-fake) than to the false positives (non-fake records delegated fake ones), one will show signs of improvement execution as for the fake class (Weiss, 2004). Another conceivable approach, utilized as a part of this stage, is to join and apply diverse calculations to perform better outcomes, by exploiting the quality of every one (Phua, 2004). Those two diverse stage's methodologies end with a standard grouping model, where there is a requirement for them to be assessed and the execution of the classifier to be measured. This is impossible as far as prescient exactness where all blunders costs the same in light of the fact that, as we have seen some time recently, on the grounds that the minority class is the most huge one (Yue, 2007) (Yang, 2006). Figure 3 indicates twenty occasions of a dataset, in which fifteen typical and five deceitful exchanges, where it is hard to isolate the fake from the no fraudulent ones. Thus, one of the ways is apply an essential grouping strategy to this dataset. Here, it is hard to assemble a decent model. In a more terrible case circumstance, one would arrange a model with a 75% of accuracy, in that it would group all occurrences as typical. This is the primary issue when managing unequal information since it gives back a grouping model with a high accuracy rate. One of the worries here would be that these conditions can't generally recognize any deceitful case.

The last imperative technique that can be connected to distinguish extortion is exceptions recognition. An exception is a perception that goes amiss such a great amount from differentperceptions as to stir doubt that it was created by an alternate mechanism (Hawkins, 2002). This implies an exception is an information record that has property estimations that an altogether different from most of alternate records. This circumstance prompts to the assurance that the qualities were not actually started. Therefore, the ways to deal with distinguish exceptions are focused on the disclosure of examples that happen in an occasionally path in the information, instead of the most conventional systems of information mining that have the objective to discover designs that happen every now and again in the information (Koufakou, 2007). Anomalies are all the time regarded as mistakes that should be evacuated amid the information cleaning period of KDD, all together for the predetermined calculation to succeed. For instance, this would be for one to think about a sensor that measures hourly air temperature in a building. Consistently, amid a 24-hour time frame, the temperature sways in the vicinity of 18ºC and 22ºC, yet at a specific hour, the temperature is recorded to be at 40ºC. This esteem would need to be considered as an anomaly since it digresses, fundamentally, from alternate or historic records. Thusly, to have proficient model that are produced from an information source, it is important to evacuate all exception record.

In any case, exceptions can once in a while prompt to the disclosure of imperative data. In the particular instance of misrepresentation recognition, anomalies can be outstanding mistakes or the can, indeed, be an indication of a fake movement. Without a doubt, utilizing the anomaly procedures, one can, likewise, distinguish misrepresentation, which is another method for battling it. Figure 4 displays the different techniques for extortion identification that is portrayed in prior record. Adjusting the dataset as observed some time recently, in a genuine setting, the common circulation of information is not the most sufficient arrangement in the application including classifiers. For this situation, one of the conceivable arrangements is to change the information, keeping in mind the end goal to make a more adjusted dataset. To achieve this undertaking, there are three methodologies that might be connected to adjust information. These procedures are over-examining, under-inspecting or both. Re-testing the dataset is a strategy that makes a more adjusted dataset by wiping out a portion of the records of the lion's share class (under-inspecting) or by creating more records of the minority class (oversampling) (Han, 2005).

Despite the fact that that condition can bring better outcomes, it additionally carries numerous more issues with it. By killing examples of the greater part class, the issue of over-fitting can happen. An over fitted model is a model excessively particular that can have a higher exactness rate in a particular dataset in any case, when it is utilized with another arrangement of information, won't deliver any helpful outcomes. That is a typical issue happening in prescient models. Expelling a portion of the occasions of the larger part class, one can make essential data misfortune, as found in the case, the prescient model that groups all information with estimations of incomes littler than 3 and estimations of costs greater than 2 as deceitful information has a prescient rate of 87% however a similar model connected to the first dataset has just a prescient rate of 70%. As observed here, it is justifiable that when under sampling happens, the greater part class, by haphazardly expelling tests, can encounter lost a portion of the valuable data (Chawla, 2003).

This is a fundamental and basic answer for adjust the dataset however as additionally under-inspecting by disposal, it can prompts to over-fitting. Another issue that can happen in this circumstance is while applying this procedure keeping in mind the end goal to expand the span of the dataset, which can prompt to a tremendous issue, when managing officially vast and lopsided datasets (Chawla, 2004). In spite of the fact that re-testing the information is a smart thought to start a more adjusted dataset, keeping in mind the end goal to help the classifiers fabricate a superior prescient model, this is a straightforward undertaking since it can make different issues that will disturb the nature of this model. Consequently, the essential procedures for demonstrating are not proficient and, subsequently, different systems must be tended to. One of these principle methods, which serve to manage over testing, is known as the Synthetic minority over-inspecting strategy (SMOTE),destroyed over-specimens the minority class by bringing into the condition manufactured cases, along the line of sections that join the k, minority class and closest neighbors (k-NN). Contingent on the measure of over-inspecting required, neighbors from the k–NN are self-assertively picked (Chawla, 2002). Utilizing SMOTE, the inductive learners can extend the minority class choice districts without prompting to the over fitting issue (He, 2005) (Pelayo, 2007). Engineered tests are produced by taking the contrast between the element vector (test) under thought and its closest neighbor, increasing this distinction by an arbitrary number (in the vicinity of 0 and 1) and adding it to the element utilized element vector. This will bring about the determination of an irregular point, along the line fragment between two particular components. This approach causes, in a powerful way, the choice area of the minority class to wind up distinctly more broad. Applying this testing strategy, since it doesn't utilize the arbitrary duplication of tests, does not make a more particular minority class, rather, it causes the classifier to fabricate a bigger choice locales, which contains close-by minority class focuses (Chawla, 2002).

3. SOCIAL NETWORK DATA MINING:

Social data mining is a new and challenging aspect of data mining. It is a fast-growing research area, in which connections among and interactions between individuals are analyzed to understand innovation, collective decision making, problem solving, and how the structure of organizations and social networks impacts these processes. Social data mining includes various tasks such as the discovery of communities, searching for multimedia data (images, video, etc.), localization, personalization, and search methods for social activities (find friends or connections), text mining for blogs or other forums. Social data mining finds several applications; for instance, in e-commerce (recommender systems), in multimedia searching (high volumes of digital photos, videos, audio recordings), in bibliometrics (publication patterns) and in homeland security (terrorist networks).

Social data mining systems enable people to share opinions and obtain a benefit from each other's experience. These systems do this by mining and redistributing information from computational records of social activity such as Usenet messages, system usage history, citations, and hyperlinks among others. Two general questions for evaluating such systems are:

(1) Is the extracted information valuable?

(2) Do interfaces based on extracted information improve user tasks performance?

Social data mining approaches seek analogous situations in the computational world. Researchers look for situations where groups of people are producing computational records (such as documents, Usenet messages, or web sites and links) as part of their normal activity. Potentially useful information implicit in these records is identified, computational techniques to harvest and aggregate the information are invented, and visualization techniques to present the results are designed. Figure shows a traditional Data mining process.Thus, computation discovers and makes explicit the "paths through the woods" created by particular user communities. And, unlike ratings-based collaborative filtering systems, social data mining systems do not require users to engage in any new activity; rather, they seek to exploit user preference information implicit in records of existing activity. The "history-enriched digital objects" line of work was a seminal effort in this approach. It began from the observation that objects in the real world accumulate wear over the history of their use, and that this wear — such as the path through the woods or the dog-eared pages in a paperback book or the smudges on certain recipes in a cookbook — informs future usage. Edit Wear and Read Wear were terms used to describe computational analogies of these phenomena. Statistics such as time spent reading various parts of a document, counts of spreadsheet cell recalculations, and menu selections were captured. These statistics were then used to modify the appearance ofdocuments and other interface objects in accordance with prior use. For example, scrollbars were annotated with horizontal lines of differing length and color to represent amount of editing (or reading) by various users.

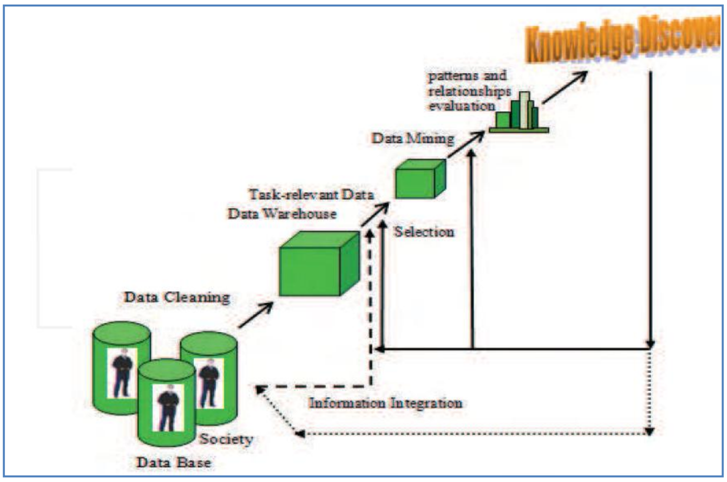

Fig.1. A traditional Social Data Mining process

The examples above mentioned are activities to which we are exposed and that without knowing we make use of the Data Mining, Due to this reason in the last years, Data Mining have had great advances in artificial intelligence in order to offer a better support to user task. Web communities have risen rapidly in recent years with benefits for different types of users. For individuals, the web community helps the users in finding friends of similar interests, providing timely help and allowing them to share interests with each other. For commercial advertisers, they can exploit the web community to find out what the users are interested on, in order to focus their targets. It would be straightforward to discover the web community if we had the detailed and up-to-date profiles of the relations among web users. However, it is not easy to obtain and maintain the profiles manually. Therefore, the automatic approaches in mining users' relationship are badly needed.

Social network describes a group of social entities and the pattern of inter-relationships among them. What the relationship means varies, from those of social nature, such as values, visions, ideas, financial exchange, friendship, dislike, conflict, trade, kinship or friendship among people, to that of transactional nature, such as trading relationship between countries. Despite the variability in semantics, social networks share a common structure in which social entities, generically termed actors, are inter-linked through units of relationships between a pair of actors known as: tie, link, or pair. By considering as nodes and ties as edges, social network can be represented as a graph. A social network is a social structure made of nodes (which are generally individuals or organizations) that are tied by one or more specific types of interdependency.

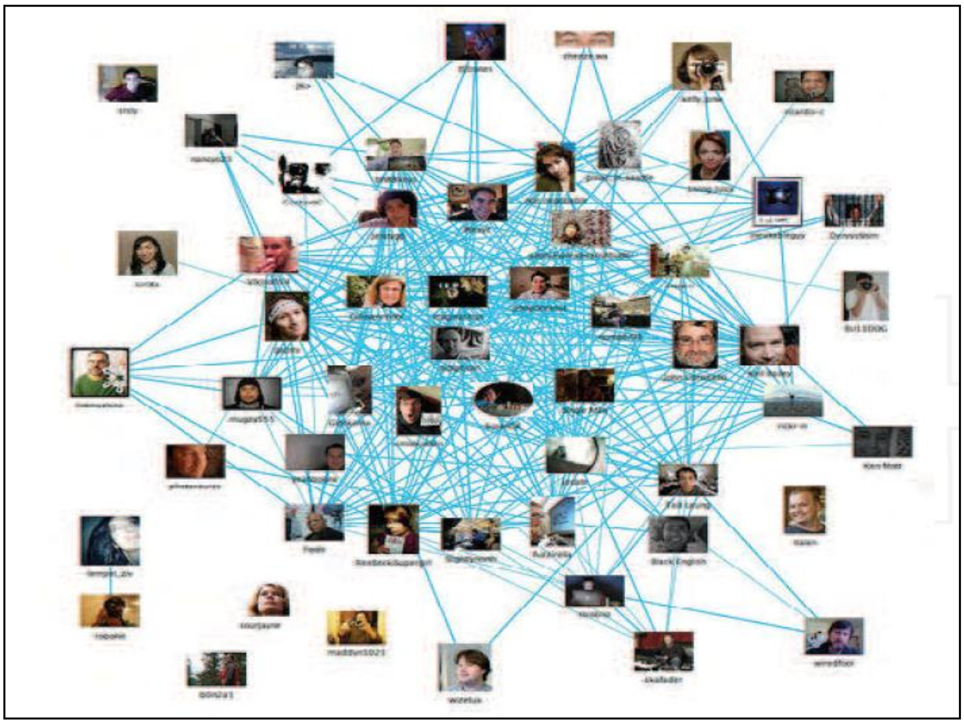

Fig. 2: Social Network Diagram

4. IDENTITY FRAUD ON SOCIAL NETWORK SERVICES:

Many people use Social Networking Services (SNSs) like daily, and link a lot of personal and sensitive information to their SNS accounts. The information generally includes friend lists, feeds from friends, non-public posts/photos, private interactions with acquaintances (such as chats and messages), and purchased apps/items. The obvious value of such information makes SNS accounts one of the most targeted online resources by hackers. SNS sites have made significant efforts to prevent identity fraud and protect users' privacy.

For example, Facebook records the regular IP addresses and devices used by each account. If an unusual IP address or device is used to log in to an account, the user is asked to answer some secret questions or enter a security code sent to the account owner's mobile device in order to verify if the login is authentic. Facebook also allows users to report account theft manually if they suspect their accounts have been compromised.

Despite all the efforts to prevent identity fraud, user privacy can be compromised by another form of breach called in-situ identity fraud—unauthorized, stealthy use of SNS accounts by attackers using the same device and network connection as the account owners. Different from other formsof identify fraud, anyone can perform in-situ identity fraud without any technology hacks. For example, anxious parents may use their children's SNS accounts to spy on the children's social status/updates or husbands/wives may check their spouses' SNS accounts if they suspect infidelity. Similarly, colleagues, supervisors, friends, and siblings, may use acquaintances' accounts for different reasons when there is a chance.

In-situ identity fraud is widespread for a number of reasons. First, people tend to choose "yes" when the browsers on their own computers ask if they want to save their (SNS) passwords for automatic logins in the future. This is especially true when people use their mobile devices because inputting passwords is inconvenient. Mobile devices make in-situ identity fraud easy in other ways, as they can be physically accessed by acquaintances or strangers, and most of them are not locked by PINs. In addition, many SNS sites use cookies to avoid the need for account authentication within a short period of time. For example, once logged into Facebook, a user does not need to log in again for up to 60 days. Given the above drawbacks, if someone (usually an acquaintance) can access an SNS user's computer or mobile device, it is unlikely that he will need a technical background to obtain the information associated with the SNS account.

DETECTION SCHEME - SNS services are not simply places for people to maintain their friend lists. They are more like platforms where people can engage various social activities, such as posting details of their own status, reading other users' comments on the news, chatting, and meeting new people. Some studies suggest that there is no typical user behavior pattern on a complicated, open platform like Facebook, as every user seems to behave differently on an SNS service. For example, some people use SNSs to satisfy their desire for self-promotion, so they spend most of their time sharing thelatest information about their status and posting the latest photos/events. On the other hand, some people may want to make new friends online, chat with old friends, or spend time discovering new social games; and some may want to stalk certain other users.

In the context of in-situ identity fraud, an SNS user can be classified as fulfilling one of the following roles:

1) Anowner which means the person uses his own account

2) Anacquaintance (as a stalker), who uses the account of someone he knows or

3) Astranger (as a stalker), who uses the account of a person he does not know.

Intuitively, when owners check their Facebook newsfeeds, they should focus more on the latest information posted by friends and use the "like" or "share" function to interact with others. By contrast, when a stalker (either an acquaintance or a stranger) browses a newsfeed, he may be more interested in old information about the stalker and/or the account holder. Also, the stalker generally does not interact with others to avoid discovery by the account holder about the identity fraud. In summary, we believe that users' behavior varies in different roles for the following reasons:

- The way people treat familiar information (and information from close friends) would be different than the way they treat unfamiliar information. Social influence is a key factor.
- People in different roles have different intentions.
- To avoid detection by the account owners, stalkers tend to not make any interaction with others. Also, they may have limited time to commit in-situ identity fraud so their browsing behavior would be even more different.

We define the above differences in SNS users' browsing behavior as role-driven behavioral diversity, which serves the rationale behind the proposed detection scheme. After a user logs in with a stored credential or existing authentication cookies, following actions happen:

Step 1: the SNS server monitors and records the user's actions for an observation period of n minutes, where n is a configurable parameter

Step 2: At the end of the observation period, the server extracts the features of the monitored session based on the recorded actions

Step 3: It then feeds the session features which characterize users browsing behavior, into a classification model

Step 4:which determines if the session owner is suspicious by predicting the label of the session?

Step 5: If the predicted label is “stalker,” the SNS server can challenge the user by asking secret questions or via a second channel, such as the account owner's mobile phone

Step 6: Alternatively, the server can implement a more sophisticated, but costly, detection scheme.

CONCLUSION

There are a few procedures and strategies that can be sent to help with extortion discovery. The unequal dataset display an issue. By the by, there are two conceivable answers for handle the circumstance, including the presentation of fake examples at a pre-preparing stage and the changing of blunder expenses at a handling level. Other than these, it is conceivable to apply exceptions as recognition strategies to signal deceitful records that digress from non-fake ones. The work exhibited here presents the use of the referenced methods to a genuine work dataset and the correlation of the outcomes acquired. Every one of the trials and tests led demonstrate that when classifiers are utilized inside lopsided datasets they are poor markers, yet when connected alongside satisfactory and particular strategies to manage this issue, the outcomes are, actually, moved forward. Other than that, another issue is concentrated here: the utilization of informal organizations to help enhance the extortion identification classifiers. This done, remembering that association and individuals are associated amongst themselves and that a few sorts of misrepresentation are propagated not by people alone but rather likewise by a relationship of these, when proposed the utilization of data concerning the interpersonal organizations of associations and individuals, in aiding in extortion identification. This review shows that a dataset when reinforce with data concerning designs recognized in informal organizations and that are normal to false associations creates some fascinating outcomes. Subsequently, the utilization of this kind of data, for some situation, may prompt to a superior classifier and, in fact, to better outcomes in extortion identification.

REFERENCES

Kumar A., Nagadevara, V. Development of Hybrid Classification Methodology for Mining Skewed Data Sets – A Case Study of Indian Customs Data [Conferência]. - [s.l.] : IEEE, 2007. - pp. 584--591.

Ahumada H., Grinblat,G., Uzal, L., Granitto, P., Ceccatto, A. A new hybrid approach to highly imbalanced classification problems [Conferência] // Eighth International Conference on Hybrid Intelligent Systems. - [s.l.] : IEEE, 2008. - pp. 386--391.

Han H., Wang, W., Mao, B-H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning [Conferência] // ICIC 2005, Part I, LNCS 3644. - [s.l.] : Springer-Verlag Berlin Heidelberg, 2005. - pp. 876--886 .

Weiss G. Mining with Rarity: A Unifying Framework. [Conferência] // Sigkdd Explorations. - [s.l.] : ACM, 2004. - Vols. 6, Issue 1 . - pp. 7--19.

Phua C., Alahakoon, D., Lee, V. Minority Report in Fraud Detection: Classification of Skewed Data [Conferência] // Sigkdd Explorations / ed. 50--59. - [s.l.] : ACM, 2004. - Vols. 6, Issue 1.

Yan Xifeng and Han JiaweigSpan: Graph-Based Substructure Pattern Mining [Journal]. - 2002.

Yang Q., Wu, X. 10 challenging problems in data mining research. [Conferência] // International Journal of Information Technology & Decision Making . - [s.l.] : IEEE, 2006. - Vol. 5. - pp. 597--604 .

Hawkins S., He, H., Williams, G., and Baxter, R. Outlier Detection Using Replicator Neural Networks [Conferência] // Procedings of the 5º Internacional Conference Data Warehousing and Knowledge Discovery. - [s.l.] : ACM, 2002. - pp. 170-180.

Koufakou A., Ortiz, E.G., Georgiopoulos, M., Anagnostopoulos, G.C., Reynolds, K.M. A Scalable and Efficient Outlier Detection Strategy for Categorical Data [Conferência] // 19th International Conference on Tools with Artificial Intelligence. - [s.l.] : IEEE, 2007.

Han H., Wang, W., Mao, B-H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning [Conferência] // ICIC 2005, Part I, LNCS 3644. - [s.l.] : Springer-Verlag Berlin Heidelberg, 2005. - pp. 876--886 .

Chawla N. C4.5 and Imbalanced Data sets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure [Conferência] // Workshop on Learning from Imbalanced Datasets II. - USA : ICML, 2003.

Chawla N., Bowyer, K., Hall, L., Kegelmeyer, W. :SMOTE: Synthetic Minority Oversampling Technique [Jornal] // Journal of Artificial Intelligence Research 16. - [s.l.] : AI Access Foundation and Morgan Kaufmann Publishers , 2002. - pp. pp. 321--357.

Chawla N., Japkowicz, N., lcz, A. Special Issue on Learning from Imbalanced Data Sets [Conferência] // Sigkdd Explorations. - [s.l.] : ACM, 2004. - Vol. 6. - pp. 1--6.

He G.,Han, H.,Wang, W. An Over-sampling Expert System for Learning from Imbalanced Data Sets [Conferência]. - IEEE : [s.n.], 2005. - pp. 537--541.

Pelayo L., Dick, S. Applying Novel Re-sampling Strategies To Software Defect Prediction [Conferência] // Fuzzy Information Processing Society. - [s.l.] : IEEE, 2007. - pp. 69--72