Performance comparative study of machine learning algorithms for automobile insurance fraud detection

Conference Paper · October 2019

DOI: 10.1109/ICDS47004.2019.8942277

CITATIONS
29

READS
450

4 authors, including:

Itri Bouzgarne
Ecole Normale Supérieure de l'Enseignement Technique
5 PUBLICATIONS **37** CITATIONS

SEE PROFILE

Mohamed Youssfi
Mohammed V University of Rabat
5 PUBLICATIONS **37** CITATIONS

SEE PROFILE

Omar Bouattane
Université Hassan II de Casablanca
289 PUBLICATIONS **1,570** CITATIONS

SEE PROFILE

Performance comparative study of machine learning algorithms for automobile insurance fraud detection

(https://ieeexplore.ieee.org/document/894227)

Bouzgarne Itri
SSDIA Laboratory
ENSET Mohammedia
University Hassan 2
Casablanca, Morocco

Youssfi Mohamed
SSDIA Laboratory
ENSET Mohammedia
University Hassan 2
Casablanca, Morocco

Qbadou Mohammed
SSDIA Laboratory
ENSET Mohammedia
University Hassan 2
Casablanca, Morocco

Bouattane Omar
SSDIA Laboratory
ENSET Mohammedia
University Hassan 2
Casablanca, Morocco

***Abstract*— Nowadays, fraud is a major problem facing the insurance industry that Big Data and Machine Learning are trying to solve. This paper deals with the evaluation of the effectiveness and the verifiability of the best-known machine learning algorithms for fraud prediction. We adopted the supervised method applied to automobile data claims of an anonymous insurance company. We aim to propose an approach that improves the relevance of the results of artificial intelligence. The study has demonstrated that RandomForest works better among all algorithms compared.**

Keywords— Machine learning, classification, insurance fraud detection, Supervised learning, Random Forest, Weka.

I. INTRODUCTION

Insurance fraud is a common phenomenon committed against insurance companies. Terry Allen, a statistician of the Medicaid Fraud Office in Utah [1], estimated that 10% of the 800 million claims for annual compensation could be stolen (Allen, 2000) [2]. On the other hand, they estimate that around 21% to 36% of claims are suspect, while the proportion of claims that are prosecuted represents only 3% of suspected cases of fraud [3,4]. In 2014, the Association of British Insurers (ABI) reported an 18% increase in claims over the previous year (Cutting corners, 2015) [5]. According to the Insurance Fraud Bureau of Australia (IFBA), the cost of fraudulent claims incurred by the industry is more than $2 billion annually, representing 10% of reported claims. The most recent study estimated that general insurance fraud in Australia costs more than $700 million annually. To deal with this scourge, the prevention process is triggered since the insurance underwriting against the false declarations, until the declaration or the compensation of a fraudulent suspected fraud. Reimbursement of claims remains the most expensive and common type of insurance fraud.

In our case study, we will focus on fraud in automobile claims, including the detection of suspicious behavior, whose objective is to learn the known patterns of fraud through machine learning, the study data is made up of insurance claims of an insurance company labeled fraudulent or not. A comparative analysis of ten machine-learning algorithms is presented in this work and we used two evaluation methods (F-Score and K-Score) as well as the “Root Mean Squared Log Error” indicator in order to judge the relevance of the results.

II. RELATED WORK

The problem of Fraud detection appears extensively in many machine learning research, and our work is placed in the vast field for fraud detection in financial system, including automobile insurance. Thus, in this paragraph, we quote some of the main research from this field of study.

Fraud detection has been covered by Stolfo et al. (1997a, 1997b) [6], developing meta-learning techniques for learning fraudulent credit card models by combining several classifiers to improve performance. In 2000, Stolfo et al. [7] proposed Cost-based metrics to train and evaluate the performance of learning systems for fraud detection in financial information systems. Artis et al, (2002) [8] demonstrated the performance of binary choice models for fraud detection of Automobile Insurance Fraud and implements models for misclassification assumptions in logistic regression mode. Phua et al. (2004) [9] proposed a meta-learning approach by hybridizing bagging and stacking together, comparing too approach for the fraud detection training, meta-learning approach against simpling approach. Pathak et al. (2005) [10] used the fuzzy logic-based logic system to find such types of frauds in insurance claims settlement. Pinquet et al. (2007) [11] suggested a statistical approach for fraud risk models, developing a two-equation model that applied to a real dataset of Spanish insurance company. Bermúdez et al. (2008) [12] Suggested the naive Bayes dichotomous model with asymmetric or skewed logit link to detect fraud from the Spanish insurance market database. In 2011, Xu et al. [13] proposed a random rough subspace based neural network ensemble for insurance fraud detection, and used rough set reduction to improve the consistency in the Insurance Company datasets. Tao et al. (2012) [14] projected Insurance Fraud Identification Research Based on a Fuzzy Support Vector Machine with Dual Membership, for "overlap" problems in insurance fraud samples. Benard and Vanduffel (2014) [15] studied mean-variance optimal portfolios and suggested method to maximize the measure balancing risk (Sharpe ratios), they demonstrated how results can be used in fraud detection. Sundarkumar and Ravi (2015) [16] proposed an hybrid approach for rectifying the data imbalance problem by employing k Reverse Nearest Neighborhood and one class

978-1-7281-0003-6/19/$31.00 ©2019 IEEE

support vector machine (OCSVM), in order to improve the performance of classifiers for automobile Insurance Fraud detection dataset. Nian et al. (2016) [17] proposed an unsupervised spectral ranking method for detection anomaly (SRA) of forged instances in fraud detection problem, using auto insurance claim data; They demonstrate that proposed SRA is much more efficient and surpasses existing outlier-based fraud detection methods. S.Subudhi and S.Panigrahi, (2017) [18] proposed a comparative study between various classifiers, applying Genetic Algorithm based fuzzy C-Means clustering for automobile insurance fraud detection.

III. Proposed Methodology

We applied the following classification techniques to datasets using the WEKA API : J48, Naive Bayes, Random Forest, Multilayer Perceptron, Machine Vector Support with Sequential Minimal Optimization, Logistic, Partial Decision Trees (PART), DecisionTable, Stochastic Gradient Descent and Adaptive Boosting. To discover which model is performing better, we need to train and evaluate the models using the most commonly used method : K-Fold cross-validation with 10 folds (Refaeilzadeh et al., 2009) [19].

The approach adopted in this work is based on the supervised learning method according to the following process:

Fig 1. The steps of the proposed approach.

1) Data Collection

The data set of our study is represented by data on insurance claims of vehicles provided by the Angoss KnowledgeSeeker Software popularly known as ‘‘carclaims.txt”. The dataset includes 15420 claims from January 1994 to December 1996, with 32 predictor variables and one target variable represents "Fraud" and "No Fraud"), the dataset have 14,497 genuine samples (94%) and 923 fraud instances (6%), with an average of 430 requests per month. In addition, the dataset has 6 numeric entities and 25 Alphanumeric features, each instance contains 33 features as described below:

- Insured personal data sheet (age, gender, marital status, etc.)
- Insurance contract details (policy type, vehicle category, deductible insurance payments, number of supplements, agent type, insurance coverages, etc).
- Accident circumstances (date of accident, accident area, policy report filed, witness presented, fault liability, etc)
- Other data of the insured (number of cars, previous claims, driver rating, etc)
- Fraud found (yes or no), feature to be predicted.

Various research was done on the same data set; The following table compares research results according to the following performance indicators: Accuracy, Recall, and Specificity.

Table 1: Comparative Performance Analysis algorithm for Automobile Insurance Fraud Detection using carclaims.txt.

Research Articles	Performance Metrics (in %)		
	Accuracy	Sensitivity / recall	Specificity
Xue et al. (2010)	88.70	-	-
Sundarkumar et al. (2015)	58.92	95.52	56.58
Sundarkumar and Ravi (2015)	60.31	90.79	58.69
Nian et al. (2016)	-	91.00	52.00
Sharmila and Panigrahi. (2017)	87.02	83.21	88.45

2) Data Pre-processing

The brute Data Set may contain anomalies or incorrect values that compromise the quality of the dataset, so preprocessing procedures may improve the performance of machine learning techniques by applying some processing tasks to prepare the input data in the best possible way.

The data set of our study is in CSV format, after importing it, we found 9 numeric type features and 24 nominal type features. When importing data, Weka applies systematically some heuristic rules to guess the probable type of each feature, althoug the heuristic cannot always guess the real type of the feature. Furthermore, some types (As Numeric) are not supported by classification models and certain Class such as Resample (Exception: UnsupportedAttributeTypeException), they require a data set with nominal features. Thus, we converted all the features to nominal just after importing it.

3) Feature Selection

Feature Selection is an effective way to solve the problem by eliminating redundant, irrelevant, or missing data, reducing computational time and improving the accuracy of learning.

In our case study, we have proposed to eliminate the features considered irrelevant, namely policy number, affiliation number, accident month, week of accident month, day of accident week, day of claim week, week of claim month and claim month. The improvement observed is illustrated in the paragraph “Experimental analysis”.

4) Classification

By comparing negative instances (instances of fraud) with their positive counterparts, the percentage of negatives represents only 6% of the total, so the result of the learning algorithms will not be optimal if we adopt the classic Cross Validation approach. Nevertheless, we can obtain a better result by rebalancing the Data set to readjust the percentage by giving the learning algorithm the data set while providing a comparable share of positive and negative examples.

For this purpose, we implemented "K-Fold cross validation" manually [20] without using crossValidateModel standard method of the weka Evaluation class. Indeed, we divide the dataset into 10 roughly equal parts; in each iteration, one of the k parts is used as a test set, while the other fold (k-1) is used for training. To split dataset into folds, we use the StratifiedRemoveFolds filter, which maintains the class

distribution within the folds. Meanwhile, to fill the number of the negative files, we proceed to the re-balancing of the data through the method "Resample" of Weka before applying testing and learning in each iteration, the results obtained after re-balancing data are improved.

5) Evaluation

The efficiency and performance of the algorithms are calculated from various metrics, the most prominent are recall (sensitivity), specificity, and precision. An efficient classifier should have higher values for these indicators. To facilitate the interpretation of the algorithm performance, Van Rijsberjen, (1979) [21] has created a synthetic measure F1-Measure or F-score, defined as the harmonic mean of the precision (True Positive Accuracy) and recall (True Positive Rate) of a binary decision rule. The F-measure recognized as the most commonly used measure to assess test's accuracy, applied to statistical analysis, machine learning, natural language processing and information retrieval.

$$F - Measure = \frac{\left((1+\beta^2) \times Precision \times Recall\right)}{\left((\beta^2 \times Precision) + Recall\right)} \quad (1)$$

$$\beta^2 = 1$$

Nevertheless, according to research of (Nakache and Metais, 2005) [22], the F-Measure has certain shortcomings, they introduce a meta-measure "K-Measure" which is an overall of F-measure and break even point (Joachims, 1998) [23].

$$K - Measure = \frac{(1+\beta^2) \times (Precision \times Recall)^{\alpha}}{\left((\beta^2 \times Precision) + Recall\right)} \quad (2)$$

$$\beta^2 = 1, \alpha = 0{,}5$$

Thus, we will integrate the two measures in our study to evaluate the relevance of the algorithms and proceed to classify them according to the outcome of each measure separately.

To support the results of our study between the two measures mentioned, we introduce another indicator : Root Mean Squared Log error (RMSE) (also called the root mean square deviation, RMSD) is commonly used to calculate square difference between the prediction and target for each point, then average those values into the root (Bouthevillain and Mathis, 1995) [24]. The higher this value, the worse the model is.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (3)$$

Where y_i is the actual expected output and $\hat{y}_i$ is the model's prediction.

IV. Experimental Analysis

1) Simulation tools

For the experimental validation, we used the Weka simulator. WEKA is an acronym for Waikato Environment for Knowledge Analysis, developed at the University of Waikato in New Zealand. It is open-source programming library written in Java with wide support for machine learning algorithms, able to solve a wide variety of automatic training tasks, such as data preprocessing, clustering, classification, regression and implementation.

2) Experimental results

In this paper, we have proposed a comparative analysis between the different classification models according to the two performance measures F-Measure and K-Measure. The RMSE indicator will also be calculated and compared with the result of the two measures; In the "Features Selection" Section, we reduced the number of features from 26 to 19 to keep only the relevant features. As presented in Table 2, before reduction we can see that Adaboost.M1 shows the best classification score (96.68%) using K-Measure, but Decision Table takes the lead (24.7%) using F-Measure.

However, after reducing the features, the results of F-Measure and K-measure are visibly improved as shown in Table 3. According to K-Measure, Random Forest keeps the best score with 99.4% (+4%). Unlike F-Measure, Decision Table keeps almost the same score 24.7% and gives up the first place to MultilayerPerceptron with 25.6% (+1.2%).

Admittedly, the two rankings are different or even opposite, indeed RandomForest is ranked last with F-Measure, whereas it is ranked first with K-Measure. In order to judge the relevance of the two measures, we introduced the RMSE indicator, as presented in Fig.2, the result obtained shows a strong rank correlation between K-measure and RMSE, indeed the lowest value is attributed to Random Forest with 29.8%. However MultilayerPerceptron is ranked 5th with 41.26%.

Table 2: Comparative Performance Analysis before Feature Selection on Dataset rebalancing:

Classifier	Recall	Precision	F-measure	K-measure
AdaBoostM1	12,67%	20,96%	15,72%	**96,68%**
RandomForest	13,32%	22,28%	16,62%	96,44%
PART	34,78%	15,90%	21,79%	92,78%
J48	45,61%	14,29%	21,76%	85,17%
MultilayerPerceptron	63,81%	15,26%	24,40%	79,05%
DecisionTable	75,83%	14,90%	**24,79%**	74,12%
SGD	76,44%	13,57%	22,91%	71,95%
Logistic	79,74%	13,39%	22,93%	70,17%
NaiveBayes	84,61%	12,27%	21,44%	66,51%
SVM	91,76%	12,64%	22,22%	65,24%

Table 3: Comparative Performance Analysis after Feature Selection on Dataset rebalancing:

Classifier	Recall	Precision	F-Measure	K - Measure
Random Forest	23,83%	19,66%	21,52%	99,46%
AdaBoostM1	28,38%	17,92%	21,94%	97,34%
PART	35,64%	15,51%	21,60%	91,89%
J48	45,61%	14,86%	22,40%	86,07%
MultilayerPerceptron	65,12%	16,10%	25,66%	79,92%
Decision Table	75,83%	14,90%	24,79%	74,12%
Logistic	84,50%	13,40%	23,14%	68,75%
SGD	88,52%	12,93%	22,56%	66,69%

NaiveBayes	87,21%	12,30%	21,55%	65,81%
SVM	91,76%	12,64%	22,22%	65,24%

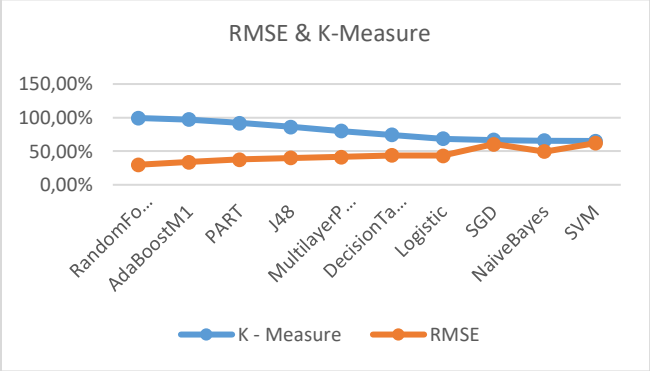

Fig 2: Correlation between K-Measure and RMSE.

V. Conclusion

In this work, we present a comparison of the ten most frequently used machine-learning algorithms and compare their performance with two evaluation methods to determine which is the most appropriate, applied to on real world data for fraud prediction.

The study shows that the Random Forest algorithm has the best performance with K-measure evaluation and the best score with RMSE. Radom Forest also has the best ratio between recall and precision, the share of fraud discovered among all frauds is 23.8%. On the other hand, we have the precision is 19.7% holding the highest score between models, this means that in 2/10 of the cases when a claim is predicted as fraud, the model is correct. Furthermore, if we compare accuracy value with the result found by the previous researches mentioned in Table 1, we have earned the highest score with 89.57 attributed to Random Forest. Moreover, we can also conclude that the performance evaluation method for a classification model based on the (Van Rijsberjen, 1979) approach using its F-Measure formula is outdated, the method should be upgraded or replaced with the (Nakache & Metas, 2005) approach.

In real life, accusing an insured of fraud is a grave act, the client risks not being indemnified for a probability of a claim suspected fraudulent, such probability depending on the result and performance of the classification model chosen. RandomForest has the best score in our case study, but the insurance companies are more prudent, as they can only open a litigation process for the piece of evidence and not on a probability basis. Otherwise, this can compromise the insurer's image and reputation, or even generate other indirect costs.

ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 777720.

References

[1] Richard J. Bolton and David J. Statistical Fraud Detection: A Review . Hand Statistical Science Vol. 17, No. 3 (Aug., 2002), pp. 235-249.

[2] Allen, T. (2000), A day in the life of a Medicaid fraud statistician. STATS 29, 20- 22.

[3] Tennyson Sharon, Salsas-Forn Pau. Claims auditing in automobile insurance: fraud detection and deterrence objectives. Journal of Risk & Insurance, Vol. 69, pp. 289-308, 2002.

[4] A. Derrig Richard, Insurance fraud, J Risk Insur, 69 (2002), pp. 271-287.

[5] Cutting corners, August 2015. Cutting corners to get cheaper motor insurance backfiring on thousands of motorists warns the ABI.

[6] Stolfo, S.J., Prodromidis, A.L., Tselepis, S., Lee, W., Fan, D.W., 1997a. JAM: Java agents for meta-learning over distributed databases. AAAI Workshop on AI Approaches to Fraud Detection. In: Proceedings of the 3rd International Conference Knowledge Discovery and Data Mining, pp. 74–81.

[7] Stolfo, S.J., Fan, D.W., Lee, W., Prodromidis, A.L., Chan, P., 2000. Cost-based modeling for fraud and intrusion detection: results from the JAM project. In: Proceedings of the DARPA Information Survivability Conference and Exposition (DISCEX'2000), vol. 2, pp. 130–144.

[8] Artis, M., M. Ayuso, and M. Guillen, 2002, Detection of Automobile Insurance Fraud With Discrete Choice Models and Misclassified Claims, Journal of Risk and Insurance, Vol. 69,pp. 325-340.

[9] Phua, C., Alahakoon, D., Lee, V., 2004. Minority report in fraud detection: classification of skewed data. Acm Sigkdd Explor. Newslett. 6 (1), 50–59.

[10] Pathak, J., Vidyarthi, N., Summers, S.L., 2005. A fuzzy-based algorithm for auditors to detect elements of fraud in settled insurance claims. Managerial Auditing J. 20 (6), 632–644.

[11] J. Pinquet, M. Ayuso, and M. Guillen, "Selection bias and auditing policies for insurance claims," Journal of Risk and Insurance, vol. 74, pp. 425-440, 2007.

[12] Bermúdez, L., Pérez, J., Ayuso, M., Gómez, E., Vázquez, F., 2008. A bayesian dichotomous model with asymmetric link for fraud in insurance. Insurance: Math. Econ. 42 (2), 779–786.

[13] Xu, W., Wang, S., Zhang, D., Yang, B., 2011. Random rough subspace based neural network ensemble for insurance fraud detection. In: Computational Sciences and Optimization (CSO), 2011 Fourth International Joint Conference on. IEEE, pp. 1276–1280.

[14] Tao, H., Zhixin, L., Xiaodong, S., 2012. Insurance fraud identification research based on fuzzy support vector machine with dual membership. In: Information Management, Innovation Management and Industrial Engineering (ICIII), 2012 International Conference on. Vol. 3. IEEE, pp. 457–460.

[15] Bernard, C. & Vanduffel, S., 2014. "Mean–variance optimal portfolios in the presence of a benchmark with applications to fraud detection," European Journal of Operational Research, Elsevier, vol. 234(2), pages 469-480.

[16] Sundarkumar, G.G., Ravi, V., 2015. A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. Eng. Appl. Artif. Intell. 37, 368–377

[17] Nian, K., Zhang, H., Tayal, A., Coleman, T., Li, Y., 2016. Auto insurance fraud detection using unsupervised spectral ranking for anomaly. J. Finance Data Sci. 2 (1), 58–75.

[18] Sharmila Subudhi, Suvasini Panigrahi, 2017. Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection ; Department of Computer Science and Engineering & IT, Veer Surendra Sai University of Technology, Burla, Odisha 768018, India.

[19] Refaeilzadeh, P., Tang, L., Liu, H., 2009. Cross-validation. In: Encyclopedia of Database Systems. Springer, pp. 532–538.

[20] Bostjan Kaluza, AshishSingh Bhatia, 2016, Machine Learning, in Java, Second Edition. pp.173-174.

[21] C. J. V. Rijsbergen, 1979. Information Retrieval. Butterworth-Heinemann, Newton, MA, USA, 2nd edition.

[22] Didier Nakache, Elisabeth Metais. Evaluation : nouvelle approche avec juges. INFORSID'05 XXIII e congrès, Grenoble, Jan 2005, X, France. pp.555-570.

[23] T.Joachims, 1998. "Text categorization with support vector machines", Learning with many relevant features. Proceedings of the 10th European Conference on Machine learning (ECML'98).

[24] Karine Bouthevillain, Alexandre Mathis, 1995. Prévisions : mesures, erreurs et principaux résultats. Economie et Statistique / Année 1995 / 285-286 pp. 89-100.