Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

A Data Mining based Model for Identifying of Spurious Behaviour in Water Utilization

Mr. Manu Y. M[1], Chaithra M. P [2], Hindushree A. S[3], Jyothi B. G[4], Rachana J. P[5]
[1]Asst. Professor, [2345]U,G Students
Department of Computer Science and Engineering
BGS Institute of Technology, B.G Nagar, Mandya-571448

***Abstract*:- Spurious(dishonourable) behaviour in drinking water utilization is a compelling problem facing water supplying companies and agencies. This behaviour results in a enormous loss of income and forms the highest percentage of non-technical loss. Finding competent analysis for identifying spurious activities has been an active research area in recent years. Intelligent data mining techniques can service water supplying companies to detect these spurious activities to decrease such losses. This analysis explores the use of two distribution techniques (SVM and KNN) to identify suspicious cheat water customers. The main encouragement of this analysis is to benefit Yarmouk Water Company (YWC) in Irbid city of Jordan to affect its profit loss. The SVM based approach uses customer load profile characteristics to display anomalous behaviour that is well-known to be interacted with non-technical loss activities. The data has been collected from the ancient data of the company billing system. The efficiency of the developed model hit a rate of over 74% which is better than the present standard prediction steps taken by the YWC. To expand the model, a opinion tool has been built using the develop model. The system will help the company to predict careful water customers to be authorized on site.**

I. INTRODUCTION

Water is an fundamental element for the uses of households, industry, and agriculture. Jordan, as several other countries in the world, undergo from water drought, which poses a threat that would affect all sectors that depend on the availability of water for the sustainability of actions for their development and accomplishment. According to Jordan ministry of water and irrigation, this issue always has been one of the biggest barriers to the economic growth and development for Jordan. This disaster situation has been aggravated by a population increase that has doubled in the last two decades. Efforts of the ministry of Water and irrigation to improve water and sanitation services are faced by managerial, technical and financial determinants and the limited amount of renewable freshwater resources. To address these challenges, Jordan ministry of water and irrigation as in many other countries is striving, through the adoption of a long-term plan, to improve services provided to citizens through restructuring and rehabilitation of networks, reducing the non-revenue water rates, providing new sources and maximizing the efficient use of available sources. At the same time, the Ministry continues its efforts to regulate the water usage and to detect the loss of supplied water. Water supplying companies incur significant losses due to fraud operations in water consumption. The customers who tamper their water meter readings to avoid or reduce billing amount is called a fraud customer. In practice, there are two types of water loss: the first is called technical loss (TL) which is related to problems in the production system, the transmission of water through the network (i.e., leakage), and the network washout. The second type is called the non-technical loss (NTL) which is the amount of delivered water to customers but not billed, resulting in loss of revenue. The management of the Yarmouk Water Company (Jordan) has a significant concern to reduce its profit losses, especially those derived from NTLs, which are estimated over 35% in the whole service area in the year 2012. One major part of NLT is customer's fraudulent activities; the commercial department manages the detection processes with the absence of an intelligent computerized system where the current process is costly, not effective nor efficient. NTL is a serious problem facing Yarmouk Water Company (YWC). In 2012 the NTL reached over 35%, ranging from 31% to 61 according to districts, which results in a loss of 13 million dollars per year. Currently, YWC follows random inspections for customers, the proposed model in this paper provides a valuable tool to help YWC teams to detect theft customers, which will reduce the NTL and raise profit. Literature has abundant research for Non-Technical Loss (NTL) in electricity fraud detection, but rare researches have been conducted for the water consumption sector. This paper focuses on customer's historical data which are selected from the YWC billing system. The main objective of this work is to use some well-known data mining techniques named Support Vector Machines (SVM) and K-Nearest Neighbor (KNN) to build a suitable model to detect suspicious fraudulent customers, depending on their historical water metered consumptions.

II. RELATED WORK

This section reviews some of the applications of data mining classification techniques in fraud detection in different areas such as Detection of Fraudulent Financial Statement, Fraud Detection in Mobile Communication Networks, Detecting Credit Card Fraud, and Fraud Detection in Medical Claims. For example, Kirkos et al. proposed a model for detecting fraud in financial statements, where three data mining classifiers were used, and namely Decision Tree, Neural network and Bayesian Belief Network.

Shahine et al. introduced a model for credit card fraud detection; they used decision tree and support vector machines SVM. In addition, Panigrahi et al. proposed a

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

model for credit card fraud detection using a rule-based filter, Bayesian classifier, and Dempsters-Shafer adder.
Carneiro et al. developed and deployed a fraud detection system in a large e-tail merchant. They explored the combination of manual and automatic classification and compared different machine learning methods. Ortega et al. proposed a fraud detection system for Medical claims using data mining methods. The proposed system uses multilayer perceptron neural networks (MLP). The researchers showed that the model was able to detect 75 fraud cases per month.
Kusaksizoglu et al. introduced a model for detecting fraud in mobile communication networks. The results showed that the Neural Networks methods MLP and SMO found to give best results. In addition, CHEN et al. proposed and developed an integrated platform for fraud analysis and detection based on real time messaging communications in social media. More recently, Syeda et al. have proposed the use of parallel coarse-grained neural networks for speeding up the data mining and knowledge discovery process. Maes et al. have outlined an automated water scam detection system by ANN as well as Bayesian belief networks (BBN). They show that BBN gives better outcomes related to scam detection and the training period is faster whereas the actual detection process is substantially faster with ANN. The neural network formed methods are, in general, fast but not so authentic. Re-training the neural networks is also a major bottleneck since the training time is quite high. Chen et al. propose a novel method in which an online test is used to collect test-responded transaction (QRT) data of users. A support vector machine (SVM) is competent with this data and the QRT models are used to predict new transactions.
Chen et al. have recently presented a personalized access for water scam detection that employs both SVM and ANN. It tries to prevent scam for users even without any transaction data. However, these systems are not fully computerized and depend on the user's expertise level. Some prober have applied data mining for water scam detection. Chan et al. divide a large set of transactions into smaller subsets and then apply distributed data mining for building models of user behaviour. The resultant base models are then combined to generate a meta-classifier for improving detection accuracy.
Brause et al. have analyze the possibility of combining advanced data mining techniques and neural networks to obtain high fraud coverage along with a low false alarm rate. Use of data mining is also complicated in the work by Chiu and Tsai. They consider web services for data interchange among banks. A fraud pattern mining (FPM) algorithm has been evolve for mining scam association rules which give information regarding the features that exist in scam transactions. Banks build up their original scam identify systems by using the new scam patterns to prevent attacks. While data mining designs are approximately authentic, they are inherently slow.
The customers who are recorded in the municipality of Gaza as water theft have been manually marked with the label 'YES' in the field scam Status, and the rest were labeled as 'NO.' The data was filtered to remove difference and noise cases. The data is normalized using z-score to fit the SVM model. Similar to all fraudulent cases, the data classes are unbalanced. SVM has a parameter set that can be used to the weight and balance the two classes. The ratio for each class was calculated to compute the parameter values. The values were multiplied by 100 to achieve a suitable ratio weight for SVM. The random subsampling was applied to the samples with class label "NO" to weight the classes for KNN and ANN. The results showed that SVM classifier has the best accuracy over the other two classifiers for the balanced samples either with consumption feature alone or with all selected features, The season consumption dataset was the most suitable for monthly and yearly datasets because it takes into consideration the seasonal consumption changes where the intelligent model raised the hit rate from 10% random inspection to 80%. This research showed that unbalanced samples gave the best accuracy for all classifiers.

A. The Support Vector Machines (SVM)

Classifier Support vector machine (SVM) is a supervised classification method. It works well for linear and non-linear data, and usable for numeric prediction in addition to classification. SVM has been widely used in different applications (i.e., object recognition, speaker identification, and hand-written digit recognition). While SVM training is relatively slow, it is highly accurate, and the problem of over fitting is less in comparison to other classifiers. SVM works by separating the training data by a hyper plane, due to the difficulty to separate the data in its original dimension. SVM uses non-linear mapping of the data into a higher dimension which enables the SVM to find hyper planes that separate the data efficiently. After that, SVM searches for the best separating hyper plane. More details about this technique can be found in the literature and most data mining books .

The K-Nearest Neighbour Classifier

The K-nearest neighbour (KNN) is a type of lazy learners, in contrast to eager classifiers like Rule-Based, Decision Tree and SVM, where the classifier constructs the model when given the training set, so it becomes ready or eager to classify. Instead, KNN is a lazy learner when given a training set it does nothing and waits until the test set becomes available. It does not build a generalization; merely it stores the training tuples or instances. KNN works by comparing a given test tuple with the training tuples that are closest or similar. Therefore KNN is based on analogy. The training tuples are stored as points in the n-dimensional pattern space, in the case of unknown test tuple, KNN finds the k tuples that are closest to the test tuple, these tuples are the K-nearest neighbours; the test tuple is classified by the majority voting of its k neighbours. The similarity can be measured using several distance metrics such as the Euclidian distance.
Let two tuples x1 = (x11, x12, ..., x1n) and x2 = (x21, x22, ..., x2n). The Euclidean distance (dist) between x1 and x2 is computed using equation (1):

$$(1,2) = \sum (1\text{-}2)2n \quad i=2 \qquad (1)$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

To determine the best value for K, this can be achieved experimentally; we start with k=1 and increment by 1 repeatedly until we obtain the minimum error rate.

III. METHODOLOGY

The CRISP-DM (Cross Industry Standard Process for Data Mining) was adopted to conduct this research. The CRISPDM is an industry standard data mining methodology developed by four Companies; NCR systems engineering, DaimlerChrysler AG, SPSS Inc. and OHRA. The CRISP-DM model consists of business understanding, data understanding, data preparation, model building, model evaluation and model deployment.

Background knowledge of the water scam issue was derived from literature searches and from four years of interaction with concerned public and private organizations such as Best Aqua Company, Edward's Advance purifier. All sites for field work were selected in consultation with an advisory committee including representatives from the above organizations, and from the Water Supplying Companies. All field sites were selected in part on the basis that they were reputed to be among the very best in the industry in terms of scam control. The reason for selecting from among the best, rather than picking a broader or more representative sample, was to be able to work from current best practice, so that any instruction ultimately offered to the industry would help advance the state of the art. The sites were also selected so as to offer, as far as possible with these sites, a broad cross section of the industry. The sites examined included three Water scam Control Units, two private insurers (one large, one much smaller), and three private corporations acting as Water contractors, all three of which were among the top five Water contractors when measured in terms of total claims volume. A list of fifteen interview subject areas (summarized below) was provided in advance to each site, with a request that interview lists be constructed to include personnel knowledgeable in each area. The interviews themselves were not formally structured.

The fifteen interview subject areas were:

(1) Statistically valid sampling procedures or scientific estimation techniques in use to measure the scope and nature of existing scam problems.

(2) Managerial attitudes towards scams, Levels of scams regarded as "acceptable price of doing business".

(3) Budget for scam control operations, and the mechanisms for setting it.

(4) Scam control philosophy/strategy. Proactive v. reactive. How the goals of justice and cost containment are balanced. Tensions between processing efficiency and prudent controls, and mechanisms for resolving same. Distinction between "investigation" and "control".

(5) Sources of investigations (cases): range of detection mechanisms, and comparative effectiveness.

(6) Staffing/Backgrounds/Resources for fraud control operations.

(7) Use of technology for fraud detection. Existing/emerging/future systems and methods.

(8) Performance measurement for the fraud control operation. Metrics, methods in use.

(9) Nature of fraud threats: existing, emerging, anticipated.

(10) Advent of Electronic Claims Processing: effects on fraud and on fraud controls; experienced, and anticipated.

(11) Criteria used for case disposition, and for selection of administrative, civil or criminal action.

(12) Relationship with law enforcement and the criminal justice system: referral mechanisms, practices; formal & informal.

(13) Experience with managed care: how fraud differs under capitated systems, and effects on fraud control operations.

(14) Perceived constraints on effective control.

(15) Anticipated effects of various reform proposals. Industry trends and their consequences.

IV. SYSTEM DESIGN

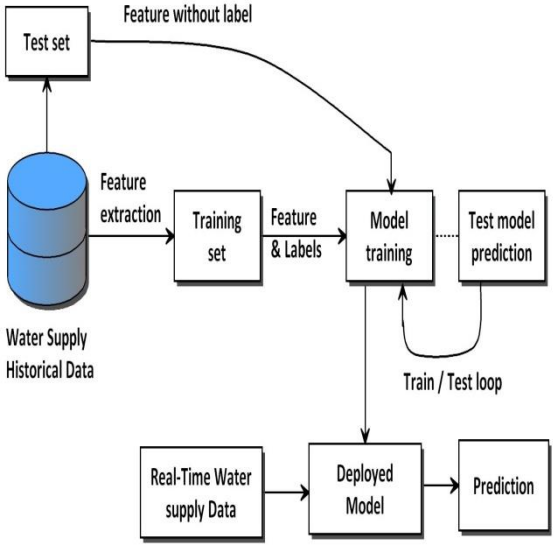

Figure : A system architecture for the detection of fraud customers.

Data mining procedures classified into two kinds as supervised and non supervised procedures. Here our model is based on the supervised modeling .“In supervised modeling, whether for the prediction of an event or for a continuous numeric outcome, the availability of a training dataset with historical data is required. Models learn from past cases”. Supervised procedures such as designation models depend on a label class must exist in the database.

The database that is Water Supply Historical Data is divided into 80 and 20 percent. Where 80% is taken as training data set and 20% taken as test data set .Training data set will go to training model. In training model we are going to apply some machine learning algorithms, like Random Forest, Naïve Bayes algorithm, Linear regression algorithm.

Training model will get the features of the training data set with label which means it indicates whether it is fraud or non-fraud and it creates a prediction model. If once the prediction model is created the testing data set given to the prediction model it detects and gives prediction to the training model without label i.e., fraud or non-fraud. So that it gives output (For example: We are giving 100 records, 80 records are matching and 20 records are not

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

matching means the accuracy is 80%. If 90 records are matching the accuracy is 90%).

V. EXPERIMENTS AND EVALUATION

This section presents the experiments, results, and evaluation. The goal of this research is detecting frauds using customers' historical data. For this purpose, model building phase is carried out using both the SVM and KNN classification techniques. For the experimentation purposes, the WEKA 3.7.10 data mining tool was used. To achieve the process of model building, 1294 customers' profile dataset was used for training the SVM and KNN models, k-fold cross-validation and holdout methods for training and testing. The model is trained using the default classifiers parameters. The SVM and KNN training and testing were done using 10-fold cross-validation and holdout methods. We analyzed the results, specifically the accuracy and a hit rate of classifying the customers into frauds or non-frauds classes. We recorded the accuracy results for each model and compared their performance in classifying new unseen tuples.

• Building the SVM Model

To build the SVM model, we used SVMLIB 1.0.6 library which is embedded within WEKA tool. We used the 10-fold cross-validation, and the holdout methods with 75%25% for training and testing respectively. In the first experiment, we used the 10-fold cross validation for training and testing, which is the default parameter of the algorithm. Table displays the confusion matrix of this model. The confusion matrix shows that SVM model scored an accuracy of 71%. This result exhibits that 920 records out of 1294 records were correctly classified, and 374 records out of 1294 records were misclassified. In the second experiment, we modified the test option to Holdout, that is, to split the input tuples into a separate training set and testing set. This is intended to measure the classifier's accuracy when changing the training and testing parameters. The parameter split percentage was set to 75%; this means that 75% of the original dataset is used for training and 25% for testing. The results accuracy was 72.4.

TABLE: SVM CONFUSION MATRIX, USING CROSS VALIDATION

		Predicted		Accuracy %
		Fraud	Not Fraud	
Actual	Fraud	394	293	61
	Not Fraud	121	526	81
	Accuracy %	76.5	67.5	71.1

VI. CONCLUSION

In this research, we applied the data mining distribution techniques for the reason of identifying customers' with scam behaviour in water utilization. We used SVM and KNN classifiers to frame distribution models for identifying suspicious scam customers. The data used in this research refer the data was collected from Yarmouk Water Company (YWC) for Qasabat Irbid ROU customers, the data covers five years customers' water utilization with 1.5 million customer historical records for 90 thousand customers. This aspect took a noticeable attempt and time to pre-process and setup the data to fit the SVM and KNN data mining classifiers. The handled observations displayed that a good achievement of Support Vector Machines and K-Nearest Neighbours had been accomplished with overall efficiency around 70% for both. The model hit rate is 60%-70% which is probably better than irregular standard analysis conducted by YWC unit with hit rate around 1% in identifying scam customers. This model proposes an smart gadget that can be used by YWC to identify the scam customers and decrease their profit losses. The adviced model helps preserving time and attempt of employees of Yarmouk water by identifying billing faults and corrupted meters. With the benefit of the considered model, the water services can increase cost recovery by decreasing organizational Non-Technical Losses (NTL's) and increasing the productivity of analysis team by onsite research of suspicious scam customers.

REFERENCE

[1] C. Ramos , A. Souza , J. Papa and A. Falcao, "Fast non-technical losses identification through optimum-path forest". In Proc. of the 15th Int. Conf. Intelligent System Applications to Power Systems, 2009, pp.1-5

[2] E. Kirkos, C. Spathis and Y. Manolopoulos, "Data mining techniques for the detection of fallacious money statements", skilled Systems with Applications, 32(2007): 995–1003

[3] Smart Cities Conference (ISC2), Trento, pp. 1-4.

[4] Y. Sahin and E. Duman, "Detecting credit card fraud by decision trees and support vector machines", IMECS, 2011, Vol I, pp. 16 – 18.

[5] S. Panigrahi, A. Kundu, S. Sural and A. Majumdar, "Credit card fraud detection: a fusion approach using dempster–shafer theory and bayesian learning, information fusion", 2009, 10(4): 354–363.

[6] Ortega P., Figueroa C., and Ruz G."A Medical Claim Fraud/Abuse Detection System supported information Mining: A Case Study in Chile", In proc of DMIN, 2006

[7] C. Richardson, "A privacy protective approach to energy thieving detection in good grids", 2016 IEEE International

[8] R. Jiang, H. Tagiris, A. Lachsz. "Wavelet-based choices extraction and multiple classifiers for electricity fraud detection", In Proc.IEEE/PES Transmission and Distribution Conf. Exhibit. 2002.

[9] Approach to Detection of Tampering in Water Meters", In Procedia Computer Science, 2015, 60: pp 413-421.

[10] B. Boser, I. Guyon, and V. Vapnik, "A coaching formula for best margin classifiers," in Proceedings of the 5th Workshop on Computational Learning Theory.