TEST Engineering & Management

May - June 20200
ISSN: 0193-4120 Page No. 3219-3222

Detecting Unethical Practices in Consuming Water by Using Data Mining Based Model

A.Arunachalam[1], G.Divya[2]

[1,2]Department of Computer Science and Engineering, Saveetha School of Engineering
Saveetha Institute of Medical and Technical Sciences
me.arun170@gmail.com[1], mailtodivya16@gmail.com[2]

Article Info
Volume 83
Page Number: 3219-3222
Publication Issue:
May - June 2020

Article History
***Article Received:** 19 August 2019*
***Revised:** 27 November 2019*
***Accepted:** 29 January 2020*
***Publication:** 12 May 2020*

Abstract

In water, fraudulent behavior is a problem faced by companies that supply water. This act ends in a bulk amount of loss of income and creates chaos. In recent years, Researches are being made for efficient measurements to detect fraudulent activities. There are classification techniques such as SVM and KNN which is explore in the paper to detect mistrustful customers. The research's only purpose is assisting the Nalco water company (NWC) in Jordan city for overcoming the profit loss. It finds a rate over 74% which predicts the NWC's manual procedure. To deploy the model, we are using a decision tool to generate the model. This project will help to predict water customers who are frauds or suspicious on the site.

***Keywords:** Datasets, NTL, KNN & SVM Model*

1. Introduction

Health care to a community cannot be provided without drinking water. So, water is vital for all living things like humans, animals, and plants. A lot of cities suffer from scarcity of water which is the major threat that affects all sectors depending on water for their development and their living.

According to NWC, the problem is caused by growth of population that is doubled in past 10 years. This issue has always been the biggest barrier to stock and economic growth. This was lead the Jordan administrator of water as in many other countries for striving, to provide the services through reconstruction and improvement of network, that reduces the non-earning water rates. At the same time, the company has to detect the waterloss. WSC obtain significant losses because of fraudulent operations in consumption of water.

The person who messes with their water meter reading to curtailbilling amount is known as fraud or cheat customer. In the current scenario, water loss is two types Technical loss and non technical loss, the first one which is the Technical Loss is meant to issues of the manufacturing system, the transferring water through network that is loss of water. the second thing is Non-Technical Loss (NTL) which is a caused by fraud customer which leads revenue loss to the company. In 2012, one major part of the company is Non-technical loss. Non-Technical Loss is a serious problem facing Nalco water company (NWC) which results in 14 million dollars loss in each year.

They are following random inspections for the customers. In this proposed model, we give a valuable tool that is really helpful for the Nalco water company. So, in this project, we are looking forward to the historical data for the billing system. The important objectives are to use well-known for SVM and KNN. These two are mainly used for machine learning algorithms. SVM is shortly termed as a support vector machine which is used for regression problems. KNN is shortly termed as k-nearest neighbors that are used for classification and regressions and to detect the fraudulent customers.

2. Literature Survey

In this literature survey, there is some classification of different detection techniques using data mining such as Fraud detection in Medical claims, Fraud detection in credit card techniques and Fraud detection in the mobile communication system. Introduced the credit card fraud detection model. They used Support vector machines and Decision tree. For our project, we purposed the KNN and SVM.

Carneiro et al.[3] made and conveyed an extortion recognition in an immense e-tail vendor. They researched the blend of manual and programmed examination and contrasted them and different AI systems. Ortega et al.[4] arranged a misrepresentation identification framework for therapeutic cases utilizing information mining techniques. The proposed framework uses multilayer preceptor neural systems. The analysts exhibit that the model had the option to identify 75 blackmail cases for every month.

Kusaksizogluet al.[5] presented a model for recognizing fraud in the mobile communication system. The outcomes appeared that the Neural Network Techniques MLP and SMO found to give the best outcomes. Moreover, CHEN et al.[6] proposed and built up a coordinated stage for fraud analysis and location detection dependent on ongoing informing interchanges in online media that is social media.

Coma-Puig et al.[7] built a system that recognizes peculiar meter readings based on models manufactured using machine learning techniques with past information. Thesystem detects meter peculiar and fraudulent customer behavior (meter altering), and it is produced for an organization that gives power and gas. Richardson et al.[8] presented novel security for detecting theft of energy in smart grids. Malicious behavior is distinguished by calculating the Euclidean separation between vitality yield estimations from the establishment over a day. These distances are then clustered for identifying outliers and malicious action.

De Fariaet al.[8] displayed a utilization instance of forensics investigation procedures applied to distinguish power theft based on altered electronic devices

The accessible literature identified with recognizing the fake exercise of fraudulent activities of Non-Technical Loss in water consumption is restricted in contrast with different segments, For example, Monederoet al.[9] built up a methodology of a set of three calculations for the recognition of meter altering in the Emasesa Company (a water circulation organization in Seville).

Nagi et al. [10] [11] [12] presented techniques classifying fraud behavior of customers in electricity utilization. The method proposed consist of mixture of two classification algorithms, Genetic algorithm (GA) and support vector machine (SVM) [12], which yield a half and half model also named GA- SVM. The procedure prepared the profile of past customers' utilization to reveal abnormality made by customers of Tenaga Nasional Berhad (TNB) [13] power facility in Malaysia. After the classification, four classes were found such as tenant change, meter replacement, faulty mater, and copious house. A special framework designed to remove such customers by considering the attributes and recognizing the four categories with theft customers. This smart system hit-rate arrives at 60% which indicated that the model raised the fraud detection activities from 3% utilizing power system in the company to, a hit pace of 60% after onsite testing inspection. The customers who are all in the municipality GAZA as water theft are manually marked with the label 'YES' otherwise the rest of the persons were 'NO'. This data is normalized using the SVM model.

CRISP-DM

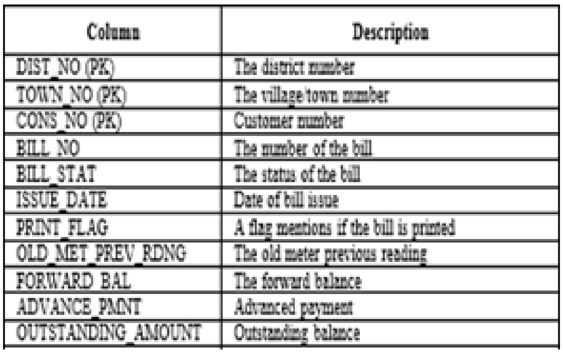
The Cross Industry Standard Process for Data Mining was embraced to lead this examination. The CRISP-DM is an industry standard information mining procedure created by four Companies;
SPSS Inc, OHRA DaimlerChrysler AG,NCR frameworks designing. The CRISP-DM model comprises of business understanding, information understanding, information readiness, model structure, model assessment, and model organization. COBOL is used for the business and financial system for companies.

Business Understanding

Nalco water organization was set up in July 2010; it is altogether claimed by the Water Authority of Jordan (WAJ). The organization built up ten branches everywhere throughout the concession district; each branch is answerable for dealing with the appropriation of water and client undertakings in its assigned zone, these branches are called Regional Organizational Units (ROU's).

Data Understanding

These areas describe the structure and nature of the gathered information. The information gathered from the charging framework which is principally utilized for giving the client's water bills. Appropriate COBOL programs have been created to extricate the most significant clients charging information into content configuration information documents. The tables are the clients' principle data table, Customers' water utilization table, and the Customers' installments table. The portrayal of the Customers' water utilizations table (connection) is introduced in Table



Column	Description
DIST_NO (PK)	The district number
TOWN_NO (PK)	The village/town number
CONS_NO (PK)	Customer number
BILL_NO	The number of the bill
BILL_STAT	The status of the bill
ISSUE_DATE	Date of bill issue
PRINT_FLAG	A flag mentions if the bill is printed
OLD_MET_PREV_RDNG	The old meter previous reading
FORWARD_BAL	The forward balance
ADVANCE_PMNT	Advanced payment
OUTSTANDING_AMOUNT	Outstanding balance

Figure 1: Consumption Table

Data preparation

This period of the knowledge discovery encourages tremendous endeavors to set up the information with greatest quality and reasonable configuration to be utilized later in the demonstrating stage.

Data Preprocessing

The utilization table of the verifiable clients' information contains around 16 million records for 109 thousand clients. It incorporates the utilization for the interim from 1990 to the present time. The clients' utilization records that are identified with ROU are around 1.5 million utilization records for around 90 thousand clients. The utilization of information for the clients is put away in a vertical arrangement.

Framework

Arrangements for the fraud detection system.

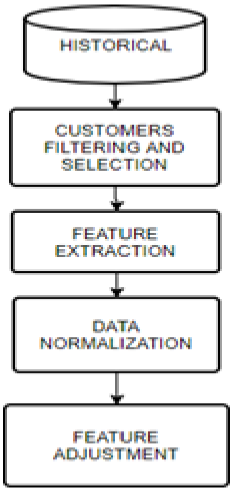


Figure 2: Data framework

Support Vector Machine

A support vector machine is a directed arrangement technique. SVM functions admirably fornon-straight direct and information, and valuable for numeric forecast notwithstanding grouping. Support vector machine has been broadly utilized in various applications (i.e., acknowledgment object, speaker distinguishing proof, and transcribed digit acknowledgment). While SVM preparing is generally moderate, it is exceptionally precise, and the issue of over fitting is less in contrast with different label.

Support vector machine works by isolating the preparation information by a hyperplane, because of the trouble to isolate the information in its unique measurement. SVM utilizes non-direct mapping of the information into a higher measurement which empowers the SVM to discover hyperplanes that different the information proficiently. From that point forward, SVM looks for the best isolating hyperplane. More structure about this technique can be establish in the literature and max data mining book.

K-Nearest Neighbor

The K closest neighbor is a sort of apathetic learner, in differentiation to anxious classifiers like Decision Tree, Rule-Based and SVM. Rather, KNN is a languid learner when given a preparation set it sits idle and holds up until the test set gets accessible. KNN works by contrasting a given test tuple and the preparation tuples that are nearest or comparable. Therefore KNN depends on similarity. The test tuple is signified by large scale voting of its k-neighbors Let two tuples.

z1 = (z11, z12, ..., z1n) and

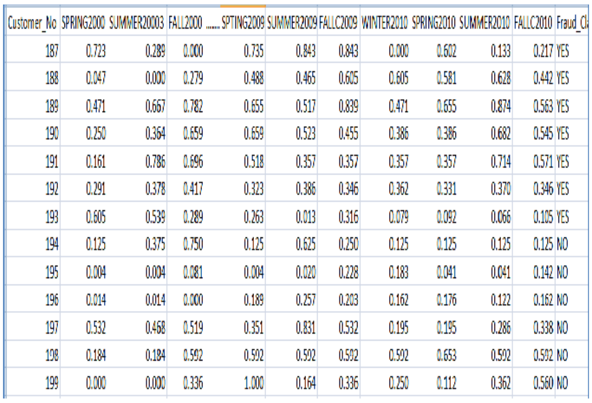
z2 = (z21, z22, ..., z2n).

dist(z1,z2) =$\sqrt[2]{\sum^{N} I=1(zi- z2i)}$

Build KNN & SVM Model

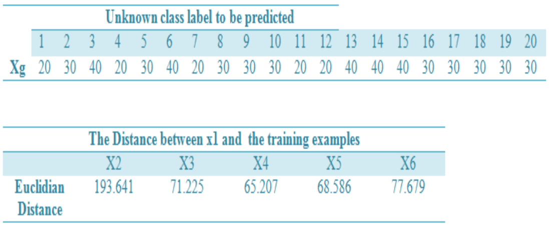
The models were constructed utilizing SVM and KNN calculations. In the dataset there are two classes of clients' profiles (Fraud and Not Fraud) the clients class names are lopsided where the known misrepresentation customers contrasted with ordinary are 2:10000 this will bring about a poor order exactness, adjusting the dataset is compulsory for the model to perform well, an irregular sub-examining is utilized in the higher class (non-extortion) to adjust the dataset, and all sifted misrepresentation clients are chosen for use in the preparation of the models.

To setup the SVM model, we run SVMLIB 1.0.6 library [28] which is installed within the WEKA tool. We used the 10-fold cross-validation, and the holdout methods with 71% - 21% for practiceand testing.



Customer_No	SPRING2000	SUMMER20003	FALL2000	.... SPTING2009	SUMMER2009	FALLC2009	WINTER2010	SPRING2010	SUMMER2010	FALLC2010	Fraud_Cl
187	0.723	0.289	0.000	0.735	0.843	0.843	0.000	0.602	0.133	0.217	YES
188	0.047	0.000	0.279	0.488	0.465	0.605	0.605	0.581	0.628	0.442	YES
189	0.471	0.667	0.782	0.655	0.517	0.839	0.471	0.655	0.874	0.563	YES
190	0.250	0.364	0.659	0.659	0.523	0.455	0.386	0.386	0.682	0.545	YES
191	0.161	0.786	0.696	0.518	0.357	0.357	0.357	0.357	0.714	0.571	YES
192	0.291	0.378	0.417	0.323	0.386	0.346	0.362	0.331	0.370	0.346	YES
193	0.605	0.539	0.289	0.263	0.013	0.316	0.079	0.092	0.066	0.105	YES
194	0.125	0.375	0.750	0.125	0.625	0.250	0.125	0.125	0.125	0.125	NO
195	0.004	0.004	0.081	0.004	0.020	0.228	0.183	0.041	0.041	0.142	NO
196	0.014	0.014	0.000	0.189	0.257	0.203	0.162	0.176	0.122	0.162	NO
197	0.532	0.468	0.519	0.351	0.831	0.532	0.195	0.195	0.286	0.338	NO
198	0.184	0.184	0.592	0.592	0.592	0.592	0.592	0.653	0.592	0.592	NO
199	0.000	0.000	0.336	1.000	0.164	0.336	0.250	0.112	0.362	0.560	NO

Figure 3: Normalization



	Unknown class label to be predicted																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Xg	20	30	40	20	30	40	20	30	30	30	20	20	40	40	40	30	30	30	30	30

	The Distance between x1 and the training examples				
	X2	X3	X4	X5	X6
Euclidian Distance	193.641	71.225	65.207	68.586	77.679

Figure 4: Class label

With the customer consumption diagram, we can calculate the fraud detection process. When choosing K=1, the nearest neighbor is z4 with the class label "Yes," therefore the model predicts that the class label for z1 is "Yes" meaning that the customer is predicted as a suspicious Fraud. Choosing K=3 the nearest neighbors are: x3, x4, x5 with class labels "Yes," "Yes," and "No" respectively, therefore (using the majority voting technique) the predicted class label for z1 is "Yes", meaning that the customer is predicted as a suspicious fraud

SVM Model

		predicted		Accuracy %
		Fraud	Not scam	
Actual	Fraud	445	202	68
	Not scam	186	461	70
	Accuracy	70	69	69.01

Figure 5: Svm Model Actual

KNN Model

		predicted		Accuracy %
		Fraud	Not scam	
Actual	Fraud	384	253	71
	Not scam	121	526	81
	Accuracy	71.5	67.5	71.11

Figure 6: KNN Model Actual

The accuracies of SVM and KNN are 70%, 71% respectively using a 10-fold cross-validation method for training and testing. SVM and KNN classifiers are a close performance in fraud detection

	Accuracy	Recall	Training & Test option
SVM	71.1%	61%	10-fold cross validation
KNN	70.0%	68%	
SVM	72.4%	68%	Holdout 75% training & 25% testing
KNN	74.3%	73%	

3. Conclusion

In this investigation, we applied the data mining classification to perceive clients' with blackmail conduct in water utilization. We utilized SVM and KNN assign to manufacture request models for identifying suspicious blackmail customers. The models were manufactured using the customers' authentic metered utilization information; the Cross-Industry Standard Process for Data Mining. The coordinated preliminaries demonstrated that well execution of Support Vector Machines and K-Nearest Neighbors (had been practiced with as a rule exactness around 70% for both.

Being used of the proposed model, the water utilities can rise cost recuperation by decreasing managerial Non-Technical Losses (NTL's) and rise the profitability assessment staff on location investigations of suspicious extortion clients.

References

[1] J. Nagi, K. Yap, S. Tiong, S. Ahmed, and A. Mohammad. "Recognition of variations from the norm and power robbery utilizing hereditary help vector machines", In Proc. IEEE TENCON Region 10 Conf., 2008, pp.1-6.

[2] N/A, "Jordan Water Sector Facts and Figures, Ministry of Water and water system of Jordan". Specialized Report. 2015.J. Han, M. Kamber, J., and Pei. Data mining: concepts and techniques, 3rd Ed, Morgan Kaufmann.2012.

[3] N. Carneiro, G. Figueira and Costa M., "An information digging based framework for charge card misrepresentation location in e-tail choice emotionally supportive networks", Decision Support Systems, 2017, 95(C): 91-101.

[4] C. Cortes and V. Vapnik, 1995. "Support-Vector Networks", Machine Learning, 1995, 20(3): 273-297Monedero I., Biscarri F.

[5] Guerrero J., Roldán M., and León C. "An Approach to Detectionof Tampering in Water Meters", In Procedia Computer Science, 2015, 60: pp413-421

[6] J. Nagi, K. Yap, S. Tiong, S. Ahmed, and A. Mohammad. "Detection of abnormalitiesand electricity theft using genetic support vector machines", In Proc. IEEE TENCON Region 10 Conf., 2008,pp.1-6.