International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 7, ISSN No. 2455-2143, Pages 241-244
Published Online November 2019 in IJEAST (http://www.ijeast.com)

SURVEY ON CLASSIFICATION ENGINE FOR MONETARY TRANSACTIONS

Ms. K. B. Satpute
Department of Computer Engineering
Sinhgad College of Engineering,
Pune, India

Abhiraj Kale
Department of Computer Engineering
Sinhgad College of Engineering,
Pune, India

Anurag Mandal
Department of Computer Engineering
Sinhgad College of Engineering,
Pune, India

Ram Krishnan
Department of Computer Engineering
Sinhgad College of Engineering,
Pune, India

***Abstract*—Tracking regular expense is a key factor to maintain a budget. People often track expenses using pen and paper method or take notes in a mobile phone or a computer. These processes of storing expense require further computations and processing for these data to be used as a trackable record. Hence, we propose a system which helps its users in tracking their expenses on an everyday basis using a web application. In this application, user can either input their bank statements or manually feed the expense. The system reads this data and classifies it using Machine Learning Algorithm based on the type of expense.**

Keywords - Machine Learning, Classification, Random Forest Algorithm, Expense Management

I. INTRODUCTION

Expense management is a necessary task and our system provides the users with a comfortable way of completing the task. We use a publicly available dataset which does not contain any personal information to train the system. The user can also manually feed the expense onto the system. The input is sent to the classification engine which then uses the custom search API to search for the listed expense. We use a supervised machine learning algorithm which needs a labelled dataset to work efficiently. The output is a clean, arranged and classified data in terms of predefined tags.

II. MOTIVATION

A. Old methods

Up until now recording or tracking expenses has been done on paper and registers, such methods are obsolete now in a world where everything is digital.

B. Data growth

This digital world includes a very high rate of data growth, hence we need to replace old methods for accounting (pen-paper) and most importantly to classify it before storing. Due to increase in digital transactions it is easier to track and maintain a record of the transaction if we have a classification engine.

III. LITERATURE REVIEW

There have been a lot of work regarding expense management that includes sharing expenses, keeping track of personal expenses in an excel sheet and some even includes Optical Character Recognition (OCR) for capturing text-based data from a picture of the bill provided by the user.

Paper 1

Chaudhuri et al.(1997) proposed a system which uses the OCR to read scripts written in two different languages i.e. Bangla and Devanagari. As the two-language had a lot of the same features because of both having the same origin, so the system could read the two languages using the same process. A set of algorithms were used for document digitization, skew detection, text line segmentation and zone separation, word and character segmentation, character grouping into basic modifier and compound character category. The limitation noted down in this study was that the efficiency was compromised and it affected the working of the system in large scale applications. We overcome this limitation by using an effective mechanism that initially performs a local search in the database for the transaction before using the search API. This helps in saving processing time when the transactions are repeated.

Paper 2

Angshuman Paul and Dipti Prasad Mukherjee et al.(2018) in their paper “Improved Random Forest for Classification”, proposed a system for improved random forest classifier that performs classification with minimum number of trees. They found out that the system had some advantages like high accuracy, handles missing values and no overfitting or underfitting. They also noted some down sides of the algorithm like it tends to be a time-consuming process and was a little complicated and difficult to interpret sometimes. In this paper they used a vaguely labelled dataset due to which the system faced difficulties for processing. We have an accurately labelled dataset using search API and

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 7, ISSN No. 2455-2143, Pages 236-240
Published Online November 2019 in IJEAST (http://www.ijeast.com)

classification engine which will ease the processing of the random forest algorithm, thus decreasing complexity in the system.

Paper 3
Yuchen Qiao, Yunlu Li and Xiaotian Lv et al.(2019) in their paper The Application of Big Data Mining Prediction Based on Improved K-Means Algorithm in 2019 proposed a system where the relationship between the 14 characteristic variables of the citizen is studied. They found out that the system worked well with many advantages like better clustering effect on big data with more attributes. But it also had some shortcomings like it had low operation efficiency and low clustering accuracy. K-Means algorithm tends to have overfitting and underfitting issues, hence affecting clustering performance which is resolved in the random forest algorithm because it dynamically alters the level of search node. This helps in achieving a greater clustering efficiency.

Paper 4
Vo Duy Thanh, Vo Trung Hung, and Doan Van Ban et al.(2013) in their paper Text Classification Based on Semi-Supervised Learning proposed a system - they present their solution and results of the application of semi supervised machine learning techniques. They discovered that the classification quality is enhanced after improvement features model but the effectiveness of semi-supervised model is low. Semi-Supervised learning algorithms does not use a labelled dataset whereas our system runs on an accurately labelled dataset using search API and classification engine. We use a supervised machine learning algorithm which uses the labelled dataset for processing.

Paper 5
Zhijie Liu, Xueqiang Lv Kun and Liu Shuicai Shi et al.(2010) in their paper Study on SVM Compared with the other Text Classification Methods, proposed a study on the application of support vector machine in text categorization. They achieved a precision rate of more than 86.26%, high adaptability and provided a simple structure. But their system was vulnerable to factors like many key parameters affects accuracy. In this paper, the authors changed many factors and discovered that the system was vulnerable to massive changes. The proposed system uses a consistent, labelled dataset and does not require much changes in the attributes. So we observe a consistent performance since our dataset remains consistent.

Sr. No	Paper	Techniques used	Findings
1	Improved Random Forest for Classification	Improved random forest classifier that performs classification with minimum number of trees.	High accuracy, handles missing values and no overfitting or underfitting Tends to be a time consuming process and was a little complicated
2	The Application of Big Data Mining Prediction Based on Improved K-Means Algorithm	Relationship between the 14 characteristic variables of the citizen is studied.	Better clustering effect on big data with more attributes. Low operation efficiency and low clustering accuracy
3	Text Classification Based on Semi-Supervised Learning	They present their solution and results of the application of semi supervised machine learning techniques	Quality is enhanced after improvement features model. Effectiveness of the model is low.
4	Study on SVM Compared with the other Text Classification Methods	Study on the application of support vector machine in text categorization.	Precision rate is more than 86.26%, high adaptability, simple structure. Many key parameters affect accuracy.

IV. EQUATIONS

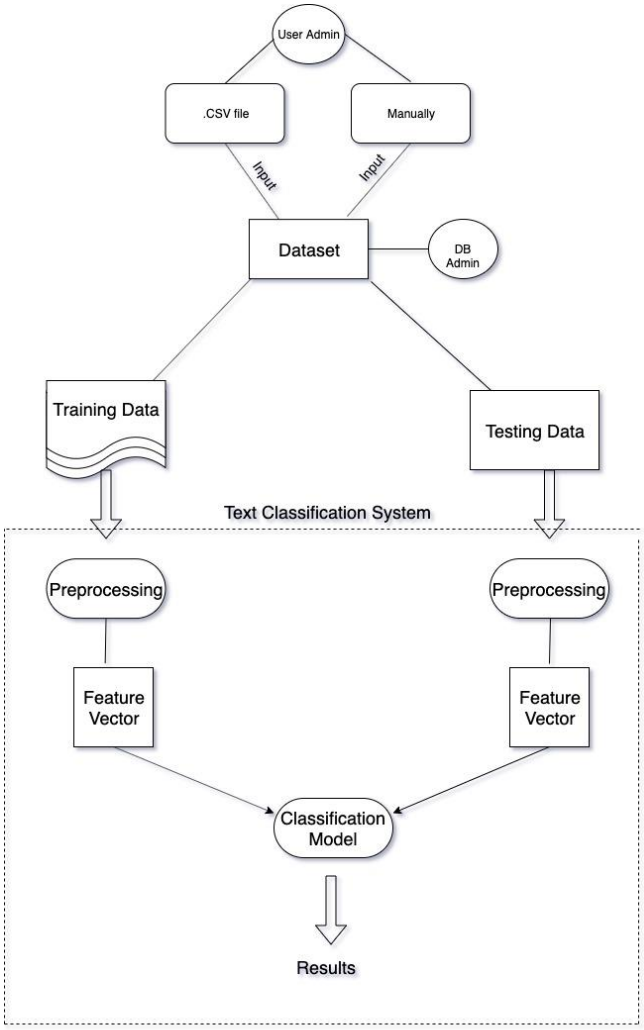
Where:

- RFfi sub(i)= the importance of feature i calculated from all trees in the Random Forest model
- normfi sub(ij)= the normalized feature importance for i in tree j
- T = total number of trees

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 7, ISSN No. 2455-2143, Pages 236-240
Published Online November 2019 in IJEAST (http://www.ijeast.com)

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T}$$

V. PROPOSED ARCHITECTURE



Architecture

Training: In this, user can give input as .CSV file or enter the transaction details manually. It will be stored in the database. In training section, the hand labelling and predefined tags will be given to the system. In testing section, some new data will be sent to the system to get the classified output. Whenever data is passed to the system it will do preprocessing on it. The process of creating features from given raw data and after this in the classification model the machine learning algorithms will do the classification and give results. We plan to explore different machine learning models like random forest, decision tree and multi-class logistic regression.

Prediction: The code of the transaction will be sent to the server and server will get more information of that code using google API's. The extracted data will be sent to the classification engine. The response will be formatted in the JSON format in the application.

VI. ALGORITHMS USED

To implement the classification engine, two algorithms will be used:

1. Rule Based Algorithm

We use Rule-based machine learning approaches to include learning classifier systems, association rule learning, artificial immune systems, and any other method that relies on a set of rules, each covering contextual knowledge.

2. Random Forest Algorithm

We use Random Forest algorithm for used for both classification and regression tasks.

VII. CONCLUSION

We foresee some issues like the model may overfit if the hand labelled dataset is too small but this can be fixed by adding a large dataset. This system has the potential to be used as a backend for a lot of banking apps. We plan to implement a feedback loop from the user. Since keyword based classification does not perform as well as other machine learning based models, the approach we took is a promising direction.

VIII. ACKNOWLEDGEMENT

Every work is a source which requires support from many people and areas. This Project has been greatly supported by the Department of Computer Engineering, Sinhgad College of Engineering. We would like to thank the Principal of Sinhgad College of Engineering Dr. S.D. Lokhande, Head of Department of Computer Engineering Prof. M.P. Wankhade and our Project Guide Prof. K. B. Satpute for their extraordinary support for this project. We would also like to appreciate the support and suggestions from our Review Committee members Prof. S. A. Joshi and Prof. N. G. Bhojne for inspiring us in this project.

IX. REFERENCES

[1] Shahed Anzarus Sabab, Sadman Saumik Islam, Md, Jewel Rana and Monir Hossain, ,2018 "eExpense: A smart approach to track everyday expense", pp. 149-156

[2] N.Zahira Jahan, M. Phil and K.I. Vinodhini, 2016, "Personalised Expense Managing assistant using Android", International Journal of Computer Techniques, pp. 5-7

[3] Sumit Yadav, Richa Malhotra and Jyoti Tripathi, 2016, "Smart Expense Management Model for Smart Homes", pp 3-5

[4] Vikas K Vijayan, Bindu K.R. and Latha Parmeswaran, 2017, "A Comprehensive Study of Text Classification Algorithms", pp. 1111

[5] Mohammad Bari, Ambaw Ambaw and Milos Doroslova, 2018, "Comparison of Machine Learning Algorithms for Raw Handwritten Digits Recognition", pp. 1-4.

[6] Angshuman Paul and Dipti Prasad Mukherjee, 2018, "Improved Random Forest for Classification", pp. 6.

[7] X.-Y. Liu, J. Wu, and Z.-H. Zhou, 2009, "Exploratory undersampling for class-imbalance learning," Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 39, no. 2. pp. 539–550.

[8] Nigel Williams, Sebastian Zander, Grenville Armitage, 2006, "A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification." , pp. 8-12.

[9] Ferdian Thung, 2013, "API Recommendation System for Software Development", pp. 897-899.

[10] Parameshwar R. Hegde, Manjunath M. Shenoy, B.H. Shekar, 2018, "Comparison of Machine Learning Algorithms for Skin Disease Classification Using Color and Texture Features", pp. 1826-1827.

[11] L. Breiman, 2001, "Random forests," Machine learning, vol. 45, no. 1, pp.5–32.

[12] Y. LeCun, 1998, "Gradient-based learning applied to document recognition," vol. 86, no. 11, pp. 2278–2324.

[13] A. Gogna and A. Majumdar, 2016, "Semi supervised autoencoder," in International Conference on Neural Information Processing. Springer, pp. 82–89.

[14] H. Ishwaran, 2014, "The effect of splitting on random forests," Machine Learning, vol. 99, no. 1, pp. 75–118.