Extraction of Bank Transaction Data and Classification using Naive Bayes

Urmika Kasi

[1]BMS College of Engineering, Department of Information Science and Engineering, Bull Temple Road, Bangalore, Karnataka, India

Abstract: ***With the increase in the number of credit card transactions, particularly over the last few years, it is important to maintain a record of the corresponding Merchant Category Codes (MCCs) of these transactions. The benefits of doing so include being able to determine interchange fee, to determine payment types for tax purposes and so on. Data mining is used to process and extract useful information such as anomalies, patterns and relationships from a large bulk of data, including large transactional data. Classification can be used to analyse such data based on their MCCs and consequently use this information for a variety of applications. This processing of data can be made efficient by transforming the data to a suitable form for analysis using pre-processing measures. In this paper, an approach is presented to extract transactional data, pre-process using pattern matching and apply a Naive Bayes classifier to perform classification based on the MCC classes of the transactions. Evaluation of the model revealed an accuracy of 0.908 and error rate of 0.092 without any majority class assumption. With a majority class assumption, the model showed a precision of 0.927, recall of 0.883 and F-Measure of 0.904. These performance measures are very good, and indicates that the consideration of Naive Bayes as classifier was an optimal choice.***

Keywords: Data mining, Classification, Naive Bayes classifier, Merchant Category Code

1. Introduction

The volumes of data being generated, particularly over the last decade, are enormous and its rate of growth is exponential. Data may be structured or unstructured, and can occur in a variety of different formats. Unstructured data constitutes nearly 90% of the digital universe, and this fact, coupled up with the numerous possible representations, makes data and text mining a task of considerable importance [1]. Data mining is the process of extracting useful information, such as anomalies, patterns and relationships from large amounts of data. It involves analysing the type of data, its quality, preprocessing it to make it suitable for application of various data mining algorithms on it, and analysing data in terms of its relationships [2]. This analysis can take the form of clustering, classification, anomaly detection and so on. Data can be categorised into many types based on the characteristic considered, such as dimensionality, sparsity or resolution. Record data indicates data to be a collection of records, which each have a fixed set of attributes. Transaction data is included in this. Data quality is a matter of particular importance, especially when considering bank transaction data, which may be very sparse. To make sure data is suitable for analysis, preprocessing algorithms are applied to it, which may involve aggregation, sampling, dimensionality reduction, feature subset selection, feature creation, discretisation and variable transformation [3]. Analysing data involves identifying relationships among data by interpreting certain metrics.

Classification is a major segment that falls under this domain, and is an analytic target function which is used to assign a category or class to some data which is taken as input, based on a model which learns from training data where each entry has a known category or membership to a class. It can be used as an explanatory tool to distinguish among various types of objects, or as a predictive measure to estimate the class/ category of unknown records. Naive Bayes classifier, based on Bayes theorem, is used when a non-deterministic relationship exists between the class variable and attribute set. It is particularly useful in cases where data is noisy, or when other external factors that affect classification.

Pattern matching is used to find some target data contained in a large set of data by exploiting the target data's characteristics which are common to all of its occurrences. Regular expressions are used to capture these patterns via a set of symbols, which when applied to a large amount of data, yield the required results if present. These expressions are constructed to represent either a deterministic or non-deterministic finite automaton [4]. Bank data such as credit card numbers, IFSC codes and so on have inherent patterns which remain static irrespective of the bank issuing the data, and hence regular expressions can be used to extract such data easily.

Banks provide transaction data to their customers most commonly in the form of periodic bank statements, which include deposits, charges, withdrawals, and the openingand closing balance for that period. This allows an analysis of possible errors, suspicious activity and expenditure. Transaction data can be used for analysis on either end- the merchant side, and customer side. Merchant Category Codes (MCCs) are assigned by a standard ISO 18245, and are used in retail financial services to control usage of corporate credit cards. They primarily reflect the category in which a merchant does business, and are also a mechanism to improve credit risk assessment [5]. They can additionally be used to infer the interchange fee paid by merchants (riskier businesses result in higher fees), to provide customers and users with points or rewards (by credit card companies) and to define card networks.

This paper presents an integrated approach to extract data from bank statements in Portable Document Format (PDF) or image formats by applying appropriate regular expressions, clean the data to make it suitable for analysis, and apply suitable algorithms to classify transaction data into basic MCCs which can then be used for numerous desired applications.

This paper is organised as follows: in section 2, an overview of background work in this area and the literature survey is presented. In section 3, the methodology used is described along with necessary equations and figures. In section 4, the experiment details and results are shown. In section 5, the conclusion along with future work on the same is given.

2. Literature Survey

Data mining and its applications in various business scenarios has proven to be extremely successful in terms of optimisation, as explained in [6].

Bayesian classifiers are widely used in a categorical context of classification, due to its superior performance. In [7]the concepts of naive Bayes classifiers, along with its situational usage, advantages and disadvantages were explored. In [8]a study was performedto understand the optimality of naive Bayes as a classifier even in the presence of strong local dependencies. The problem of extending the traditional naive Bayes model to the classification of uncertain data and demonstrates its high accuracy in such situations was addressed in [9]. In [10], a particular instance of the Naive Bayes classifier, in document classification was investigated. Naive Bayes and decision trees were appliedto bank data for detection of credit card fraud in [11]. In [12], an analysis of the misuse of MCCs in credit scoring systems using various machine learning algorithms was done. Authors in [13] used naive Bayes and decision trees on bank data for detection of credit card fraud from bank data using, and analysed its performance in comparison to C4.5, and concluded high precision and recall from the Naive Bayes classifier.

Before processing and analysing data, however, there must be efficient methods to extract data and convert it into a suitable form. The usage of pattern matching in different contexts by examining characteristic pattern matches or object pattern matches, and its application in theory-based research was explored in [14]. In [15]a comprehensive study was presented on how regexes can be used for complex information extraction and presents and analyses algorithms for the same. Authors in [16] performed feature selection and text tokenisation on strings of data for text classification, and evaluated it based on standard classification performance measures.

3. Methodology

3.1. Naive Bayes

Bayesian classifiers are used to classify data into class labels particularly when the relationship between the categories and attributes are non-deterministic. Non-deterministic relationships imply an inability to classify a class label to a set of attributes with absolute certainty despite the presence of a replica of the combination of that set of attributes previously [3]. Such classifiers assign labels to features by applying Bayes Theorem by modelling their probabilistic relationships. The formula for class conditional probability is as follows [3]:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (1)$$

where:

- X is the attribute set
- Y is the class variable
- P(Y|X) is the posterior probability
- P(Y) is the prior probability
- P(X) is the evidence
- P(X|Y) is the class conditional probability

Naive Bayes classifiers mainly require the estimation of class conditional probability, since the evidence remains constant for different values of Y, and the prior probability can easily be estimated by computing the fraction of records that belong to that particular class. This is done by adopting the class independence assumption among all attributes in X, which is as follows [3]:

$$P(X|Y=y) = \prod_{i=1}^{d} P(X_i|Y=y) \quad (2)$$

where:

- X= $\{X_1, X_{2,} \ldots, X_d\}$ where X is the attribute set and consists of *d* attributes

These classifiers are characterised by [17, 18, 19]:

- Robustness to isolated noise points and missing values
- Robustness to correlation
- Efficiency with respect to computation and prediction
- Incremental learning
- Resistance to overfitting
- Ability to handle large set of attributes

3.2. Performance Measures

Once a classification model has been built, it is important to analyse its performance in terms of accuracy of its predictions with respect to training data over the testing dataset. Classification models can face high inaccuracies due unequal distribution of instances amongst different class labels, making the training process not as comprehensive as desired. A confusion matrix is used to summarise the performance of a model accordingly to gain insight about what the model is lacking or excelling at. The rows represent the class of instances, and columns represent the corresponding predicted classes [20]. Though it can be used for multiple classes, it is generally applied with two-class concepts since the most information can be derived from this [21]. Furthermore, since analysis with a confusion matrix is necessary when there exists an asymmetry amongst classes, the more important class can be treated as “positive” (conventionally assigned as class 1), while the less important counterpart can be treated as “negative” (conventionally assigned to class 0). The matrix can be represented as below:

Table 1: Confusion Matrix

		Predicted class	
		0	1
Actual class	0	TP	FN
	1	FP	TN

where:
- TP is True Positive,
- FP is False Positive,
- FN is False Negative and
- TN is True Negative

Several inferences can be made from these values such as [22, 23]:

$$\text{Accuracy} = (TP + TN)/(TP + FP + FN + TN) \qquad (3)$$

$$\text{Misclassification error} = (FP + FN)/(TP + FP + FN + TN) \qquad (4)$$

$$\text{True Positive Rate (TPR) or Recall or Sensitivity} = TP/(TP+FN) \qquad (5)$$

$$\text{False Positive Rate (FPR)} = FP / (TN + FP) \qquad (6)$$

$$\text{Precision} = TP / (TP + FP) \qquad (7)$$

$$\text{F-Measure} = (2*\text{Recall}*\text{Precision})/(\text{Recall}+\text{Precision}) \qquad (8)$$

It is important to consider all these measures together rather than separately. For example, a classifier that is precise but inaccurate renders itself useless for all practical purposes. F-Measure is one such parameter of estimation which is derived from recall and precision, and is significant because it requires a balance among the two measures [22, 24]. It is useful in cases of uneven class distribution, which can cause accuracy alone to be an insufficient metric for model estimation.

Also, the Receiver Operating Characteristic (ROC) analysis involves the consideration of TPR and FPR to estimate sensitivity and specificity respectively such that an increase in sensitivity would result in a decrease in specificity [17]. The test is more accurate if the curve is closer to the left-hand border and the top border of the ROC area. The test is less accurate if the curve comes to the 45-degree diagonal of the ROC area.

4. Experiment and Results

This research aimed to extract transaction data and classify them into appropriate MCCs. This was done by obtaining data, converting it into a suitable format and then applying appropriate pre-processing measures to make it ready for analysis. A Naive Bayes classifier was then used for the MCC classification.

4.1. Data Collection and Preprocessing

Transactional data poses many issues. The task of obtaining sufficient data was challenging mainly due to the difficulty in obtaining bank data since they are kept confidential for privacy and security reasons. Furthermore, such data tends to be skewed, and extremely large due to the enormity in the number of card transactions that are performed on a monthly basis, making scrutiny and analysis on a large scale a problem of high computational complexity [25]. For the purpose of this research, bank statements from a number of willing people were received over many months, with details such as account number and bank balance blacked out to maintain privacy. The bank statements received covered a variety of banks such as American Express, Federal Bank, Standard Chartered Bank and Kotak Mahindra Bank to ensure the model could receive and process a variety of data patterns corresponding to different banks, rather than a single, particular one.

Since the data received was unstructured and in the form of PDFs, preprocessing was necessary to convert and compile this data accordingly to receive the transactions. Also, since different banks issue statements in different formats, it was necessary to generate generalised patterns that would retrieve data from all these statements correctly.

Firstly, the data was converted from PDF files to text files. Since several entities of bank data had fixed, recurring patterns, RegEx patterns were generated to remove and keep data for analysis accordingly. The resulting data after pattern extractions and segmentation was written back into the text file. The required data for the dataset was then taken from these text files written into a single CSV file. The dataset consists of four columns (date, desc, amount and cat) and 1160 rows. The first column, *date*, indicates the date of the transaction. Again, this could be present in different possible formats such as MM-DD-YYYY, DD-MM-YYYY, MM-YY, MM-DD and so on. The second column, *desc*, consists of the transaction description in the statement, which is used to determine the category. The third column, *amount*, indicates the monetary amount involved in the transaction, regardless of whether it is being deducted or received. The fourth column, *cat*, is the category of the transaction, which are the MCCs.

Table 2: Sample of Dataset

Date	*desc*	*Amount*	*cat*
Jun-01	BLISS CHOCOLATES INDIA BANGALORE	224	MCCs 5700-7299 (Miscellaneous Stores)
Jun-02	THE LEELA PALACE BANGALORE	13616	MCCs 3500-3999 (Lodging)
Jun-14	EMIRATES AIRLINES DUBAI	29415	MCCs 3000-3299 (Airlines)

Since the dataset was and not large enough to classify into individual MCC codes (which are nearly 4000 in number), categories were assigned based on the 13 interval classes of MCCs, which are as follows:
- MCCs 0001-1499 (Agricultural Services)
- MCCs 1500-2999 (Contracted Services)
- MCCs 4000-4799 (Transportation Services)
- MCCs 4800-4999 (Utility Services)
- MCCs 5000-5599 (Retail Outlet Services)
- MCCs 5600-5699 (Clothing Stores)
- MCCs 5700-7299 (Miscellaneous Stores)
- MCCs 7300-7999 (Business Services)
- MCCs 8000-8999 (Professional Services and Membership Organisations)
- MCCs 9000-9999 (Government Services)
- MCCs 3000-3299 (Airlines)
- MCCs 3300-3499 (Car Rental)

- MCCs 3500-3999 (Lodging)

After creating the dataset, a suitable proportion of data was set aside for training and testing. These two segments of data were then randomised individually to ensure independence and higher accuracy.

Since the training data had to manually be generated, categories were manually assigned to this data, ensuring no error. The testing data was then assigned categories based on this training data.

4.2. Results and Evaluation

Nearly 30% of the data was set aside for training. The model analysed transaction statement descriptions in the training data by stripping down text and extracting tokens from these. For example, if 'MUSICNOTE' should give 'MUSIC' and 'NOTE'. This feature extraction was used by the model to predict the category of a new transaction by analysing its features against what the model already knows.

4.2.1. Primary approach

When the dataset was considered without any biases regarding a majority class or category, an accuracy of 0.908 was seen from the model. The misclassification error was 1-0.908=0.092.

The result obtained indicates that the model performs well and is good at distinguishing between the 13 different categories. It performs well under conditions of imbalance and unequal class distribution too.

4.2.2. Secondary approach

Since the nature of the dataset indicates that "MCCs 5700-7299 (Miscellaneous Stores)" is a majority class, a second approach to classifier evaluation could be taken where 'MCCs 5700-7299 (Miscellaneous Stores)" is assigned to the positive class and all other classes are regarded as a negative class. This allows a more thorough analysis of the classifiers's behaviour since these assumptions permits the computation of additional measures such as precision, recall and F-Measure. The confusion matrix generated was:

Table 3: Confusion Matrix of Model

		Predicted class	
		0	1
Actual class	0	114	9
	1	15	68

where:

- Class 0 represents "MCCs 5700-7299 (Miscellaneous Stores)"
- Class 1 represents the remaining the classes as a singular negative class

The accuracy is (114 + 68) / (114 + 9 + 15 + 68) = 0.883.
The misclassification error is 1- 0.883 = 0.117.
The TPR or recall is (114 / 114 + 15) = 0.883.
The FPR is (9 / 9 + 68) = 0.117.
The precision is (114 / 114 + 9) = 0.927.
The F-Measure is (2 * 0.883 * 0.927) / (0.883 + 0.927) = 0.904.
The high values of F-Measure, precision and recall indicate that the model performance is very good even in the presence of a majority class and uneven class distribution.

5. Conclusion and Future Work

This paper commenced by presenting an overview of data mining and classification, and investigated the need for the same to appropriately handle and process the bulk of data being generated in today's world. Pattern matching and regular expressions were explored as one of the techniques of extract and process data. It then looked into the particular case of bank transaction data, and the applications of classification of this data into MCCs, such as categorisation and restriction of transactions by payment brands and issuers. It then proceeded into section 2 to present a background and work done by other researchers in the area. Section 3 indicated the method and approach taken by providing an in-depth analysis of the optimality of Naive Bayes as a classifier and the measures necessary to evaluate its performance. Section 4 provided details regarding the experiment, including dataset creation, data pre-processing, creation of the model, training, testing and finally evaluation. When the data was regarded without biases with respect to a majority category, an accuracy of 0.90834 was obtained, and an error of 0.09615 was seen, implying the model performed very well even when multiple class categories were considered. In a second approach to evaluation, the data was considered to be classified into two categories, where "MCCs 5700-7299 (Miscellaneous Stores)", being a majority class, was considered a positive class while the remaining categories were considered as a collective negative class. The measures obtained were: precision of 0.927, recall of 0.883 and F-Measure of 0.904.

Dataset creation was time-consuming and tedious owing to manual creation and the nature of the data being evaluated. Obtaining a variety of account statements over a long period of time proved to be difficult. Methods are being investigated to check whether the extent of automation in the processes of data collection and creation of the dataset can be furthered. By obtaining a larger variety and quantity of data, the performance of the model can be estimated with greater accuracy. Also, the existence of a majority class can lead to the model being biased towards it. This is also referred to as the class imbalance problem [23,26,27]. Future work will investigate whether techniques can be applied to alleviate this problem.

References

[1] Balas, Valentina Emilia, Neha Sharma, and Amlan Chakrabarti. "Data Management, Analytics and Innovation." Proceedings of ICDMAI 1 (2018).

[2] Kumbhar, V. S., K. S. Oza, and R. K. Kamat. Web mining: A Synergic approach resorting to classifications and clustering. River Publishers, 2016.

[3] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to data mining.* Pearson Education India, 2016.

[4] Mohri, Mehryar. "String-matching with automata." *Nord. J. Comput.* 4, no. 2 (1997): 217-231.

[5] Wang, Long. "Detection of Merchant Category Codes Application Based on Root-seeking Fast Hierarchical Clustering Algorithm." (2018).

[6] Bharati, Mrs, and M. Ramageri. "Data mining techniques and applications." (2010).

[7] Rish, Irina. "An empirical study of the naive Bayes classifier." In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, pp. 41-46. 2001.

[8] Zhang, Harry. "The optimality of naive Bayes." *AA* 1, no. 2 (2004): 3.

[9] Ren, Jiangtao, Sau Dan Lee, Xianlu Chen, Ben Kao, Reynold Cheng, and David Cheung. "Naive bayes classification of uncertain data." In *2009 Ninth IEEE International Conference on Data Mining*, pp. 944-949. IEEE, 2009.

[10] Jadon, Ekta, and Roopesh Sharma. "Data mining: document classification using Naive Bayes classifier." *International Journal of Computer Applications* 167, no. 6 (2017): 13-16.

[11] Husejinovic, Admel. "Credit card fraud detection using naive Bayesian and C4. 5 decision tree classifiers." *Husejinovic, A.(2020). Credit card fraud detection using naive Bayesian and C* 4 (2020): 1-5.

[12] Su, Chih-Hsiung, Fengjun Tu, Xinyu Zhang, Ben-Chang Shia, and Tian-Shyug Lee. "A ensemble machine learning based system for merchant credit risk detection in merchant MCC misuse." *Journal of Data Science* 17, no. 1 (2019): 81-106.

[13] Husejinovic, Admel. "Credit card fraud detection using naive Bayesian and C4. 5 decision tree classifiers." *Husejinovic, A.(2020). Credit card fraud detection using naive Bayesian and C* 4 (2020): 1-5.

[14] Trochim, William. "Outcome pattern matching and program theory." *Evaluation and program planning* 12, no. 4 (1989): 355-366.

[15] Li, Yunyao, Rajasekar Krishnamurthy, Sriram Raghavan, Shivakumar Vaithyanathan, and H. V. Jagadish. "Regular expression learning for information extraction." In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 21-30. 2008.

[16] Ikonomakis, M., Sotiris Kotsiantis, and V. Tampakas. "Text classification using machine learning techniques." *WSEAS transactions on computers* 4, no. 8 (2005): 966-974.

[17] Bužić, Dalibor, and Jasminka Dobša. "Lyrics classification using naive bayes." In 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1011-1015. IEEE, 2018.

[18] Lantz, Brett. *Machine learning with R*. Packt Publishing Ltd, 2013.

[19] Sammut, Claude, and Geoffrey I. Webb. *Encyclopedia of machine learning and data mining*. Springer Publishing Company, Incorporated, 2017.

[20] Caelen, Olivier. "A Bayesian interpretation of the confusion matrix." *Annals of Mathematics and Artificial Intelligence* 81, no. 3-4 (2017): 429-450.

[21] Cichosz, Pawel. Data mining algorithms: explained using R. John Wiley & Sons, 2014.

[22] Doreswamy, Hemanth KS. "Performance evaluation of predictive classifiers for knowledge discovery from engineering materials data sets." *arXiv preprint arXiv:1209.2501* (2012).

[23] Provost, Foster. "Machine learning from imbalanced data sets 101." In *Proceedings of the AAAI'2000 workshop on imbalanced data sets*, vol. 68, pp. 1-3. AAAI Press, 2000.

[24] Powers, David MW. "What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes." *arXiv preprint arXiv:1503.06410* (2015).

[25] Zareapoor, Masoumeh, and Pourya Shamsolmoali. "Application of credit card fraud detection: Based on bagging ensemble classifier." *Procedia computer science* 48, no. 2015 (2015): 679-685.

[26] Japkowicz, Nathalie. "Learning from imbalanced data sets: a comparison of various strategies." In *AAAI workshop on learning from imbalanced data sets*, vol. 68, pp. 10-15. 2000.

[27] Chawla, Nitesh V. "Data mining for imbalanced datasets: An overview." In *Data mining and knowledge discovery handbook*, pp. 875-886. Springer, Boston, MA, 2009.

[28] Ramasubramanian, Karthik, and Abhishek Singh. *Machine learning using R*. No. 1. New Delhi, India: Apress, 2017.

Author Profile

Urmika Kasi is a student at BMS College of Engineering and will complete the B.E course in Information Science and Engineering in 2021.