Utilizing Various Machine Learning Techniques to Classify Data in the Business Domain

Garima Malik
Student
Department of Computer Science & Engineering
Indira Gandhi Delhi Technical University for Women

Aakansha Rathore
Student
Department of Computer Science & Engineering
Indira Gandhi Delhi Technical University for Women

Sonakshi Vij
Research Scholar
Department of Computer Science & Engineering
Indira Gandhi Delhi Technical University for Women

Abstract

Machine learning techniques are commonly deployed in various real-time applications in order to generate interesting inferences, which helps in bridging the gap between relevant knowledge and the user. This paper explores the transformation of the data-sets into consolidated information. A data set that concerns the details of Indian companies, both private and government, is analyzed using K-means clustering, support vector machines and decision trees. Such an analysis will help in making the user fully informed about the budding companies and entrepreneurs in various fields that concerns them. Based upon the analysis, the company is categorized as public, private and one person company. The root node error is found to be minimal in the analysis using decision trees. The users can use this analysis to classify the company class and company status depending on factors such as company's authorized capital and paid-up capital, which further helps them to understand the revolutionized industrial environment.
Keywords: Machine Learning, K-Means clustering, Support Vector Machine, Classification, Decision Trees

I. INTRODUCTION

The industrial sector in India was initiated in India in 1855 after Industrial revolution it has grown significantly well in terms of manufacturing and industrial sector. In today's era where start-up is the new trending profession in terms of business, marketing and economic welfare. India is largely promoting various types of companies such as public, private and entrepreneur to revolutionized Indian economy [1]. The analysis done in this paper contributes towards analysing the different company's trends, class and status. Company's growth statistics can also be seen as one of the feature of the system. This paper uses various machine learning techniques to classify companies information based upon authorized and paid up capital. Integrating industrial information with supervised and unsupervised machine learning techniques [2], meaningful inferences can be carved out from the data which can be used for further analysis. The user base of this system will be primarily the company analyst or new entrepreneurs. This paper is structured as follows: section 2 shows the preliminaries corresponding to this paper, section 3 describes the proposed system, section 4 describes the results and section 5 presents the conclusion and future scope.

II. PRELIMINARIES

This section describes the various preliminaries that would be required in doing the desired analysis.

A. Machine Learning Techniques Used in the System for Classification

This sub section describes the classification techniques used in the system to classify the company's class and status based upon the authorized and paid-up capital. The techniques incorporated in the analysis are as follows:

1) K Means Clustering:

This is a classification technique which is used to classify the data in the form of clusters. It basically works on different distance measures such as Euclidean distance, Manhattan Distance and Chebychev distance[3].It tries to minimize the within group dispersion and maximize the between-group dispersion[4]. This paper classifies company class such as private public or one person.

2) Decision Trees (Conditional Inference Trees):

It is also a machine learning technique used to classify numerical and textual data[5].This technique creates tree by splitting the attributes present in the respective dataset and whichever attribute got the highest information gain will be selected to become a node. The analysis in this paper creates regression trees (also called conditional inference trees) using R libraries and classifies the company status such as Active, Dormant and Strike off.

3) Support Vector Machines:

It is also a supervised machine learning technique which is used for both regression and classification challenges [6]. It trains the data set by assigning values to each class with the help of probabilistic binary linear classifier. This analysis uses SVM to classify the company class and accuracy of this algorithm describes in the Results section.

B. Performance Measures of Machine Learning Techniques

This sub section explains the performance measures used in the paper to analyse the efficiency of machine learning algorithms.

- Withinness: It is the measure of withinness between the intra cluster distances [3].
- Betweenness: It is the measure of betweenness between the inter cluster distance [4].
- Sum of Squares By Cluster: The sum of squares by cluster is a measure of the total variance in your data set that is explained by the clustering [8].
- Complexity Parameter: The complexity parameter (cp) is used to control the size of the decision tree and to select the optimal tree size [7].

C. Economic Terms Related to this Paper

This section will describe the basic terminology used in the analysis and user base of the system can use these terms to deploy the system.

- Authorized Capital: This is the minimum amount of capital which company needs to authorize at the time of registration at ROC [14].
- Paid-up Capital: Paid-up capital is always less than the authorized capital. It is the amount of money a company has received from shareholders in exchange for shares of stock [13].
- Company Class: Registrar of Companies features 3 major classes for registered companies as Public, Private and One person.
- Company Status: Company status can be classified into Active, Dormant and Strike off.

III. ANALYSIS AND DISCUSSION

This section describes the analysis done with the concerned data set.

A. Description of Datasets and Preprocessing of Data

This sub section explains the datasets used in the system to classify the information. Initially this system considers the data from the government of India website which was raw data and statistics of that data is explained in the table 1.This system used 35 states data which includes 28 states and 7 union territories and this paper shows the results of 6 major states such as Bihar, Haryana, Manipur, Delhi, Gujarat and Punjab [9].

All the 6 datasets corresponding to these states can be referred from Table 1.After the data extraction, data cleaning is the next process to be done .All the extracted data belongs to years 19th century and onwards. And to implement machine learning techniques, there is a need of recent data so we have neglected the data before 2000 to apply algorithms on it. After data cleaning datasets attributes can be analysed through the table 1.

Table – 1
Dataset statistics of original and pre-processed data

ORIGINAL DATASET				*AFTER PREPROCESSING DATASET*			
STATES	*NO OF ROWS*	*ATTRIBUTES*	*YEARS*	*STATES*	*NO OF ROWS*	*ATTRIBUTES*	*YEARS*
BIHAR	*19,669*	*14*	*1932-14*	*BIHAR*	*13,664*	*6*	*2000-14*
MANIPUR	*397*	*14*	*1947-14*	*MANIPUR*	*307*	*6*	*2000-14*
DELHI	*2,19,750*	*14*	*1902-14*	*DELHI*	*1,96,949*	*6*	*2000-14*
GUJARAT	*81,185*	*14*	*1900-14*	*GUJARAT*	*58,869*	*6*	*2000-14*
HARYANA	*26,115*	*14*	*1901-14*	*HARYANA*	*21,651*	*6*	*2000-14*
PUNJAB	*26,167*	*14*	*1909-14*	*PUNJAB*	*18,071*	*6*	*2000-14*

B. Implementation of Proposed System

This section will describe the working prototype of the system. This system will try to solve the problems of new budding entrepreneurs who tries to analyse the no of private, public and one-person companies registered under government of India using ROC i.e. Registrar of Companies[10].This analysis is basically classification oriented which will predict the company class and company status parameters associated with the companies registered under ROC. We have used machine learning techniques in backend to classify the information and initially all the datasets are trained according to SVM classifier [6].

Mathematical equation associated with SVM trained model is as follows:

$f(x) = (u \,.\, x) + c$

$\frac{1}{n} \sum_{k=1}^{n} v(u \,.\, x_k + b, y_i) + \|u\|^2$

Subject to $\min_k$, $u.x_k = 1$

Minimize $\|u\|^2$, subject to:

$(u \cdot x_k + c) \geq 1$, if $y_k = 1$
$(u \cdot x_k + c) \leq 1$, if $y_k = -1$
The two constraints can be compacted to:
$y_k (u \cdot x_k + c) \geq 1$
Where:
X & Y are two input data sets, $x \in X$ & $y \in Y$
hyperplanes $(u.x) = 0$ where u is normalized w.r.t a set of points X such that: $\text{Min}_k |u.x_k| = 1$.
u is complexity term
v is zero-one loss function, $l(y,\bar{y}) = 1$, if $y \neq \bar{y}$ otherwise 0

K means clustering and decision trees are other two techniques used in the analysis. Both the techniques were implemented to classify the company information in a efficient manner. Linear equations are used to divide the hyperplanes on SVM plots[12].All the 6 states SVM plots presented in fig1 shows the 3 major classes a company can have and white colour denotes the private class , pink signifies the public class and fluorescent blue is the colour for one person .After that K-means clustering is shown in fig 2 in which 3 clusters of different colour is representing 3 different status of company[11].On the other hand this paper also analyses the data for 35 states such as stack plot of class wise distribution of companies so that user can see the growth of public , private and one-person companies.

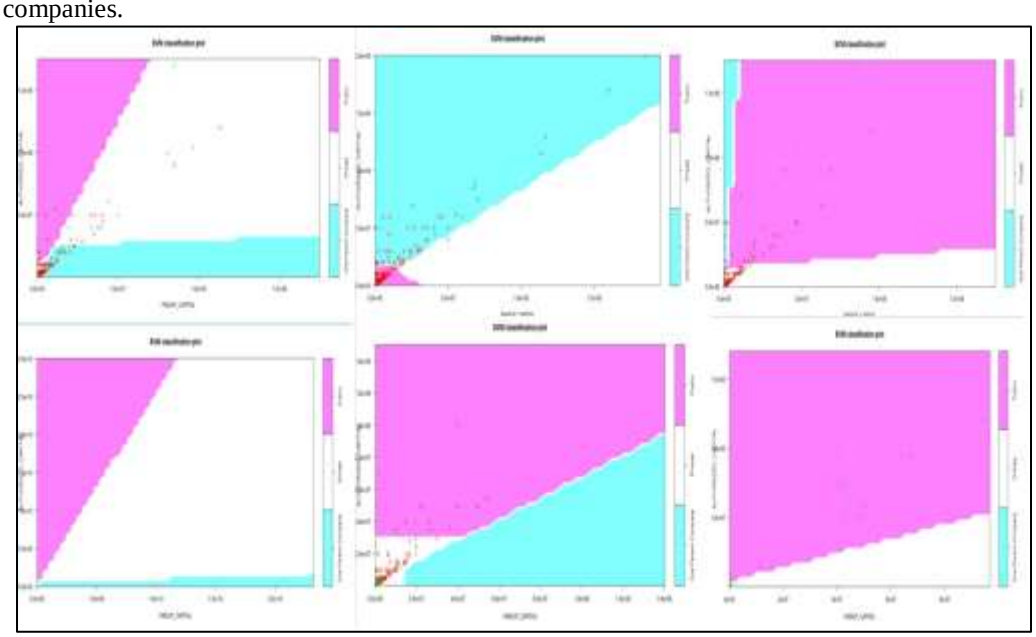

Fig. 1: Showing the support vector machine of 6 major states starting Bihar to Manipur

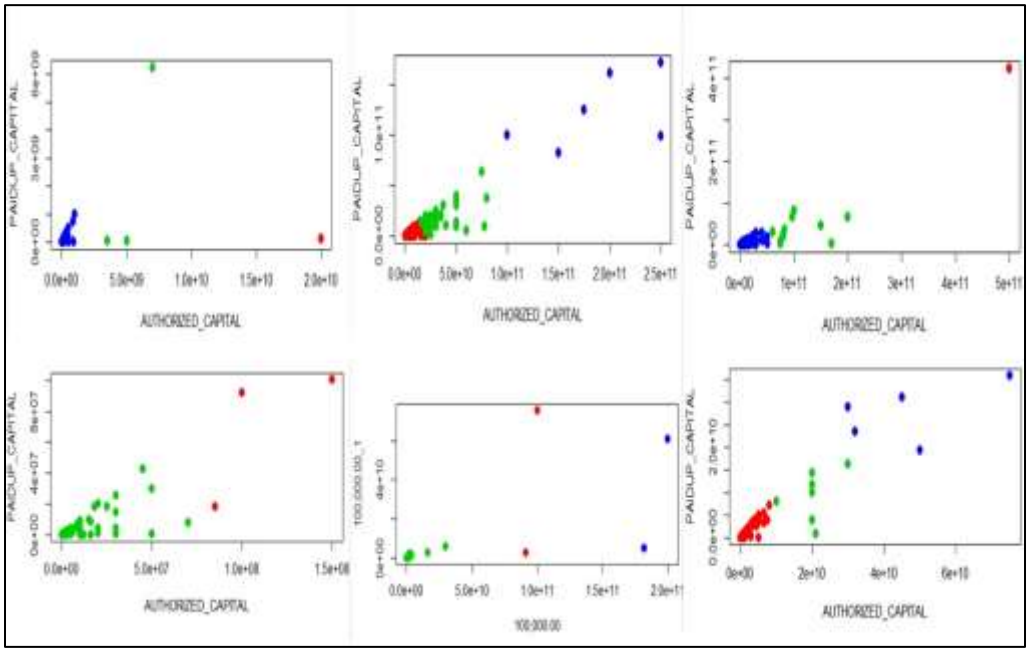

Fig. 2: Showing the cluster plot of 6 major states starting from Bihar to Manipur

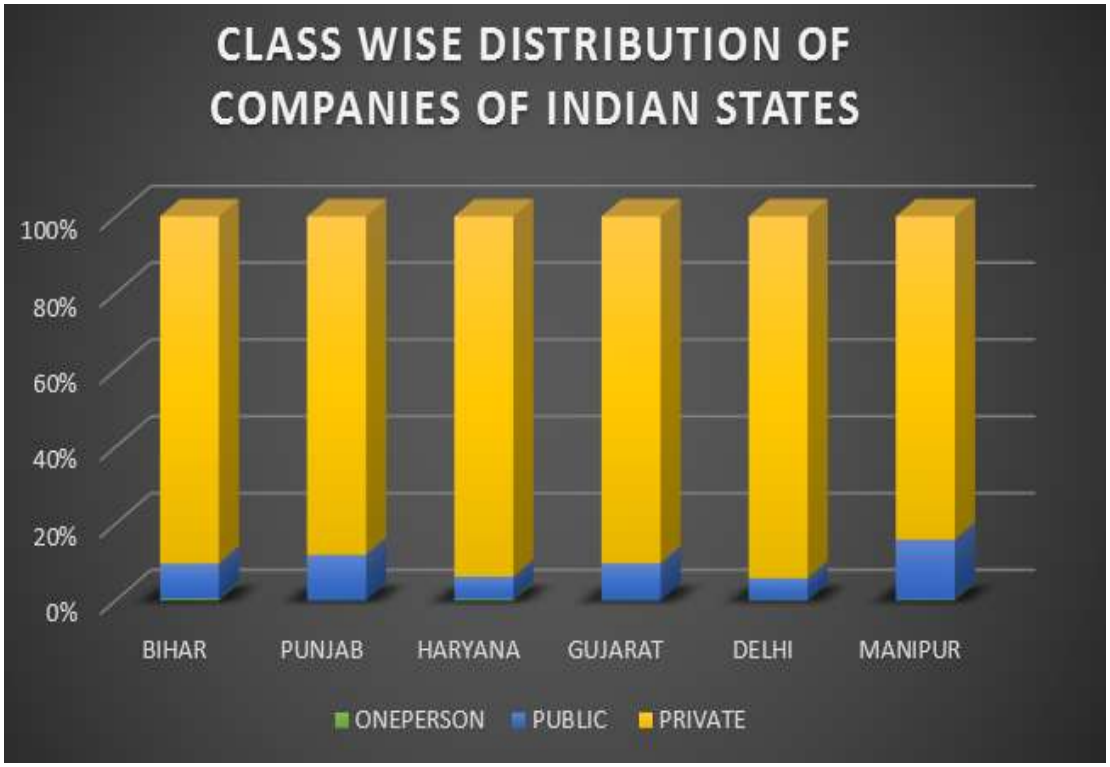

Fig. 3: Stack plot of states having public, private and one-person as company classes

IV. RESULTS AND DISCUSSIONS

This section describes the results obtained through the analysis. All the machine learning techniques have been implemented successfully on the datasets and results are presented in the form of tables in this section. Accuracy of all the techniques are shown in the tables along with that various other parameter values concerned to respective techniques are shown below. The average accuracy for the whole system is 90.75% and individually Decision tree corresponds to 90.8% and SVM accounts for 90.7%. The respective user base will provide authorized and paid up capital along with year of registration of the concerned company and our system will give the appropriate results according to the information provided by the user. Table 3 and 4 showing some instances of predictions performed by the system on the basis of authorized and paid-up capital along with that Table 2 explains the statistics related to implementation of K means clustering.

Table - 2
Performance parameters of K-means clustering algorithm.

STATES	*ITERATION BY K-MEANS CLUSTERING*	*BETWEENESS*	*WITHINNESS*	*VARIENCE IN DATA*	*NO OF CLUSTERING VECTOR*
BIHAR	*3*	*5.37647e+20*	*8.924452e+18*	*92.1%*	*3*
MANIPUR	*1*	*7.494209e+16*	*8.212029e+15*	*71.1%*	*2*
DELHI	*3*	*3.56093e+23*	*2.383987e+22*	*86.8%*	*3*
GUJARAT	*3*	*5.661514e+23*	*2.985083e+22*	*90.7%*	*3*
HARYANA	*3*	*1.801642e+22*	*3.834063e+20*	*85.7%*	*3*
PUNJAB	*3*	*9.641781e+22*	*2.671764e+21*	*94.4%*	*3*

Table – 3
Predictions related to company class using support vector machine algorithm.

STATES	*YEAR OF REGISTRATION OF COMPANY*	*AUTHORIZED CAPITAL*	*PAIDUP CAPITAL*	*OBSERVED COMPANY CLASS*	*PREDICTED COMPANY CLASS*	*ACCURACY OF SVM ALGORITHM*
BIHAR	*2014*	*1,00,000.00*	*1,00,000.00*	*PRIVATE*	*PRIVATE*	*92.1%*
MANIPUR	*2015*	*60,00,000.00*	*54,00,000.00*	*PUBLIC*	*PUBLIC*	*93.6%*
DELHI	*2016*	*2,00,000.00*	*2,00,000.00*	*PUBLIC*	*PUBLIC*	*89.3%*
GUJARAT	*2017*	*50,00,000.00*	*48,56,200.00*	*PRIVATE*	*PRIVATE*	*86.1%*
HARYANA	*2018*	*100,00,000.00*	*100,00,000.00*	*PUBLIC*	*PUBLIC*	*91.1%*
PUNJAB	*2019*	*5,00,000.00*	*1,00,000.00*	*PRIVATE*	*PRIVATE*	*92.0%*

Table – 4
Predictions of company status using Decision tree algorithm

STATES	*ACCURACY OF DECISION TREE*	*YEAR OF REGISTRATION OF COMAPNY*	*OBSERVED COMPANY STATUS*	*PREDICTED COMPANY STATUS*	*NO OF SPLITS*	*COMPLEXITY PARAMETER*	*ROOT NODE ERROR*
BIHAR	*90.4%*	*2014*	*ACTIVE*	*ACTIVE*	*8*	*0.012971*	*0.095408*
MANIPUR	*84.3%*	*2015*	*ACTIVE*	*ACTIVE*	*1*	*0.034412*	*0.15668*
DELHI	*93.8%*	*2016*	*DORMANT*	*DORMANT*	*1*	*0.000000*	*0.061104*
GUJARAT	*91.4%*	*2017*	*STRIKE OFF*	*STRIKE OFF*	*8*	*0.1216633*	*0.085483*
HARYANA	*95.1%*	*2018*	*ACTIVE*	*ACTIVE*	*8*	*0.0315896*	*0.048077*
PUNJAB	*89.8%*	*2019*	*DORMANT*	*DORMANT*	*1*	*0.000000*	*0.10171*

V. CONCLUSION AND FUTURE SCOPE

The paper analyzes the data from the business domain that concerns the details of the various private and government based companies. This paper performs classification and prediction using Machine learning techniques such as K-means clustering, support vector machine and decision trees. The classification is based on company's authorized capital and paid-up capital from which the user can easily predict the company class i.e. private, public and one person. The root node error is found to be minimal in the analysis using decision trees. In the future, such an analysis could be done using fuzzy support vector machine and fuzzy C-means clustering.

REFERENCES

[1] http://www.mapsofindia.com/my-india/india/development-in-india-after-independence
[2] Huang, G., Song, S., Gupta, J. N., & Wu, C. (2014). Semi-supervised and unsupervised extreme learning machines. IEEE Transactions on Cybernetics, 44(12), 2405-2417.
[3] Singh, A., Yadav, A., & Rana, A. (2013). K-means with Three different Distance Metrics. International Journal of Computer Applications, 67(10).
[4] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence, 24(7), 881-892.
[5] Olaru, C., & Wehenkel, L. (2003). A complete fuzzy decision tree technique. Fuzzy sets and systems, 138(2), 221-254.
[6] Hua, S., & Sun, Z. (2001). A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. Journal of molecular biology, 308(2), 397-407
[7] Magerman, D. M. (1995, June). Statistical decision-tree models for parsing. In Proceedings of the 33rd annual meeting on Association for Computational Linguistics (pp. 276-283). Association for Computational Linguistics.
[8] Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 411-423.
[9] https://data.gov.in/catalogs/sector/Industries-9323
[10] http://www.mca.gov.in/
[11] G.Malik, D.K.Tayal, A.Singh, S.Vij, Applying Data Analytics to Agricultural Sector for Upcoming Smart Cities, Proceedings of the 11th INDIACom, 4th International Conference on "Computing for Sustainable Global Development" , INDIACom 2017, In press
[12] Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. Neural processing letters, 9(3), 293-300
[13] Brimmer, Andrew F. "The setting of entrepreneurship in India." The Quarterly Journal of Economics 69.4 (1955): 553-576.
[14] Chakrabarti, R., Megginson, W., & Yadav, P. K. (2008). Corporate governance in India. Journal of Applied Corporate Finance, 20(1), 59-72.