International Journal of Mathematical, Engineering and Management Sciences
Vol. 4, No. 4, 861–869, 2019
https://dx.doi.org/10.33889/IJMEMS.2019.4.4-068

Imbalanced Ensemble Classifier for Learning from Imbalanced Business School Dataset

Tanujit Chakraborty
SQC & OR Unit
Indian Statistical Institute, Kolkata, 700108, India
E-mail- tanujit_r@isical.ac.in

(Received October 17, 2018; Accepted April 25, 2019)

Abstract

Private business schools in India face a regular problem of picking quality students for their MBA programs to achieve the desired placement percentage. Generally, such datasets are biased towards one class, i.e., imbalanced in nature. And learning from the imbalanced dataset is a difficult proposition. This paper proposes an imbalanced ensemble classifier which can handle the imbalanced nature of the dataset and achieves higher accuracy in case of the feature selection (selection of important characteristics of students) cum classification problem (prediction of placements based on the students' characteristics) for Indian business school dataset. The optimal value of an important model parameter is found. Experimental evidence is also provided using Indian business school dataset to evaluate the outstanding performance of the proposed imbalanced ensemble classifier.

Keywords- Business school problem, Imbalanced data, Hellinger distance, Ensemble classifier.

1. Introduction

Out of the many reasons behind the closing down of many of the private business schools, the foremost one is the unemployment of Master of Business Administration (MBA) students passing out of these business schools. The most challenging job for administrations is to find out the optimal set of parameters for choosing the right candidates in their MBA program which will ensure the employability of the candidates. Attracting students in business schools are highly dependent on the business schools' past placement records. If the right set of students are not selected for a few years, the number of unplaced students will certainly accumulate, resulting in the damage of reputation for the business school. One needs to develop a model in such a way that the model ensures appropriate feature selection (selection of important student's characteristics) with a decision on the optimal values or ranges of the features and higher prediction accuracy of the classifier as well. In our previous works, we proposed a hybrid classifier based on classification tree (CT) and artificial neural network (ANN) (to be referred to as hybrid CT-ANN model in the rest of the paper) to solve the business school problem (Chakraborty et al., 2018). Theoretical consistency of the hybrid CT-ANN was also studied (Chakraborty et al., 2019). In this article, we identified a vital property of the business school dataset, i.e., its imbalanced nature. Usual classifiers make a simple assumption that the classes to be distinguished should have a comparable number of instances (Wang and Alexander, 2016). Many real-world datasets including the business school dataset are skewed, in which majority portion of the cases belong to one class and fewer examples belong to the other class, yet usually more interesting class. There are also the cases where the cost of misclassifying minority examples is much higher in terms of the seriousness of the problem in hand. Due to higher weightage are given to the majority class, conventional classifiers tend to misclassify the minority class examples as the majority, results in a high false negative rate. In this particular example of business school dataset, it is clearly a two-class classification problem with the class distribution of 80:20, where a straightforward method of guessing all instances to be placed would achieve an accuracy of 80%.

There are broadly two ways to deal with imbalanced data problems. One such way to deal with the imbalanced data problems is to modify the class distributions in the training samples by applying sampling techniques. Sampling techniques include oversampling the minority class to match the size of the majority class and/or under-sampling the majority class to match the size of the minority class. Sampling is a popular approach to handle the data imbalance as it simply rebalances the data at the data pre-processing stage. But it has the obvious deficiencies like under-sampling majority instances may lose potential useful information of the dataset and oversampling increases the size of the training dataset, which may increase computational cost. Nonetheless, sampling is not the only way for handling imbalanced datasets. There exist some specially designed "imbalanced data-oriented" algorithms which perform well on unmodified original imbalanced datasets. One of the most celebrated paper in the literature is hellinger distance decision tree (HDDT) which uses hellinger distance (HD) as a decision tree splitting criterion, and it is insensitive towards the skewness of the class distribution (Cieslak et al., 2012). An immediate extension to this work is HD based random forest (HDRF) (Su et al., 2015). Another breakthrough in the literature is the class confidence proportion decision tree (CCPDT), a robust decision tree algorithm which can also handle original imbalanced datasets (Liu et al., 2010). It is to be noted that "imbalanced data-oriented" classifiers are sometimes preferred since they work with original datasets. We are therefore motivated to ask: Can we create an ensemble imbalanced data-oriented classifier which can improve the performance of HDDT, mitigate the need of sampling and solve the Indian business school data problem?

In response to this question, we proposed an ensemble classifier for feature selection cum classification problems which can be used to solve the imbalanced business school dataset problem. Our proposed ensemble classifier has the advantages of both the HDDT and ANN algorithm and performs well in high dimensional feature spaces. The optimal choice of an important model parameter is also proposed in this paper. Further numerical evidence based on business school dataset shows the robustness of the proposed algorithm.

This paper is structured as follows. In section 2, we describe the proposed ensemble model. The theoretical results are presented in section 3, and experimental analysis is shown in section 4. Section 5 is fully devoted to the concluding remarks of the paper.

2. Methodology

2.1 An Overview on HDDT

HDDT uses HD as the splitting criterion to build a decision tree. HD is used as a measure of distributional divergence and has the property of skew insensitivity (Rao, 1995). Let (Θ, λ) denote a measurable space. For any binary classification problem, let us suppose that P and Q be two continuous distributions with respect to the parameter λ having the densities p and q in a continuous space **Ω**, respectively. Define HD as follows:

$$d_H(P,Q) = \sqrt{\int_{\Omega} \left(\sqrt{p} - \sqrt{q}\right)^2 d\lambda}.$$

It is noted that HD doesn't depend on the choice of the parameter λ. The bigger the value of HD, the better is the discrimination between the features. For the application of HD as a decision tree criterion, the final formulation can be given as follows:

$$d_H(X_+, X_-) = \sqrt{\sum_{j=1}^{K} \left(\frac{|X_{+j}|}{|X_+|} - \frac{|X_{-j}|}{|X_-|} \right)^2}$$

where $|X_+|$ indicates the number of examples that belong to the majority class in training set and $|X_{+j}|$ is the subset of the training set with the majority class and the value j for the feature X. A similar explanation can be written for $|X_-|$ and $|X_{-j}|$ but for the minority class. Here K is the number of partitions of the feature space X.

2.2 An Overview on ANN

Neural network models are inspired by biological neural system. The network functions are determined mostly by the connections between neurons. The training of a neural network can be done by performing a certain function by altering the values of the connections (weights) between elements. Neural networks are trained so that inputs (feature vectors) lead to a specific target output (class level). The neural network is adjusted, based on a comparison of the output and the target, until the network output values match the predicted class. Mapping function used in ANN is very flexible. Given the appropriate set of weights, this function can approximate almost any functional form to any degree of accuracy. This function approximation is usually done by an activation function (for example, sigmoid, logsig, tansig, etc). While training the neural net with any standard dataset, the problem of overfitting can be avoided by training the network for a limited number of epochs. Feedforward backpropagation is a standard gradient descent algorithm where the weights are shifted along the negative of the gradient of the performance function (Rumelhart et al., 1986). Typically, a new testing input leads to an output that is quite similar to the correct output if it is properly trained for the input vectors used in training samples. Complex neural networks have more than one hidden layers in its architecture.

2.3 Proposed Imbalanced Ensemble Classifier

The motivation behind designing an ensemble classifier for imbalanced datasets is that one we would like to work with the original dataset without taking recourse to sampling. Here we are going to create an ensemble classifier which will utilize the power of HDDT as well as the superiority of neural networks. In the proposed imbalanced ensemble classifier (to be denoted by IEC in the rest of the paper), we first split the feature space into areas by HDDT algorithm. A set of essential features are chosen using HDDT and redundant features are removed from the dataset. We then build a neural net model using the important variables obtained through HDDT algorithm. Also, the prediction results obtained from HDDT are used as another input feature in the input layer of the ANN. The effectiveness of the proposed classifier lies in the selection of the essential features and use of predicted classes of HDDT followed by the ANN model. The inclusion of HDDT output as an additional input feature not only improves the model accuracy but also increases class separability. The informal workflow of our proposed IEC model, shown in Figure 1 is as follows:

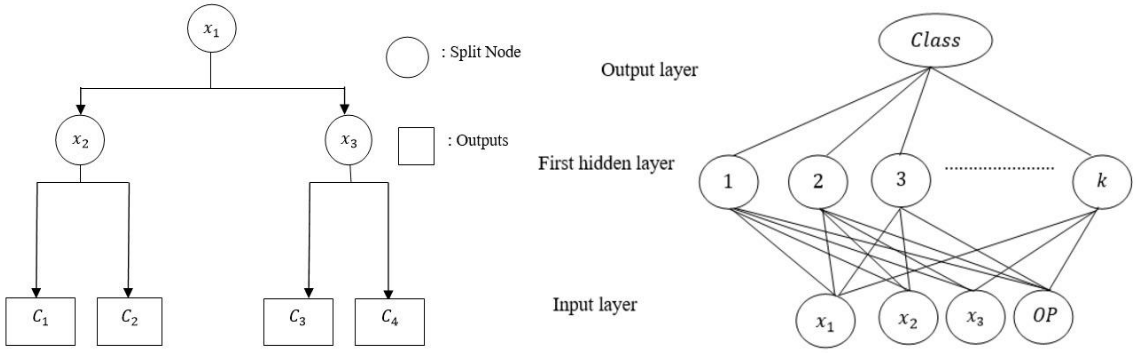

Figure 1. An example of ensemble classifier where x_1, x_2, and x_3 are important features obtained using HDDT and *OP* is HDDT output

- Sort the feature value in ascending order and find the splits between adjacent different values of the feature. We determine the binary conditional probability divergence at each split using HD measure.
- Record the highest divergence as the divergence of the whole feature. Choose the feature that has maximum HD value and grow unpruned HDDT. Using the HDDT algorithm, build a decision tree. Feature selection model generated by HDDT takes into account the imbalanced nature of the dataset.
- The predicted classes of HDDT algorithm will be used as an additional feature in the input feature space of the ANN model. Export most essential input variables obtained using HDDT along with additional feature (HDDT outputs) to the ANN model.
- Since the output results of HDDT has been incorporated as an additional feature along with other important features obtained by HDDT in the input layer of ANN, the number of hidden layer is chosen to be one.
- A one hidden layered ANN with sigmoid activation function having number of neurons in the hidden layer to be $O\left(\sqrt{\frac{n}{dlogn}}\right)$, where n is the number of training examples, d is the number of input features in the ANN model, is trained (see section 3). And finally, record the classification results.

IEC not only handles imbalance through the implementation of HDDT in selecting features but also improves the performance of the classifier by incorporating better classification results for the dataset obtained from HDDT and the model gets improved using ANN algorithm. This algorithm is a two-step pipeline approach such as handling imbalanced class distribution, selecting important features and getting an improved ensemble classifier. The optimal characteristics of students that has an impact on the placements, can be chosen by the proposed model. Also, future predictions while modelling the imbalanced dataset can also be done by IEC.

3. Optimal Value of IEC Model Parameter

Our proposed IEC has the following architecture: first, it extracts important features from the feature space using the HDDT algorithm, then it builds one hidden layered ANN model with the important features extracted using HDDT along with HDDT outputs as an additional feature. We find out the optimal choice on the number of neurons in the hidden layer of the proposed model.

Let $\underline{X}$ be the space of all possible values of p features and C be the set of all possible binary class labels. Considee a training sample with n observations, $L = \{(X_1,C_1), (X_2,C_2),..., (X_n,C_n)\}$, where $X_i = (X_{i1},X_{i2}, ...,X_{ip}) \in \underline{X}$ and $C_i \in C$. We build IEC with HDDT given features and OP as another input feature in the model. The dimension of the input layer in the ANN model, to be denoted by d_m $(\leq p)$, is the number of essential features obtained by HDDT + 1. We have used one hidden layer in the model due to the incorporation of *OP* as an input information in the model. One-hidden layered neural network yields strong universal consistency and that is evident from universal approximation theorem (Devroye et al., 2013). In IEC model, we have used one hidden layer with k neurons. Thus, the proposed IEC model becomes less complex and less time consuming while implementing the model. After elimination of redundant features by HDDT and incorporating *OP* as another input vector, let us now consider the following training sequence $\xi_n = \{(Z_1, Y_1), ..., (Z_n, Y_n)\}$ of n i.i.d copies of (Z, Y) taking values from $R^{d_m} \times C$. A classification rule realized by one-hidden layered neural network having logistic sigmoid activation function is chosen to minimize the empirical L_1 risk, where the L_1 error of a function $\psi : R^{d_m} \rightarrow \{0, 1\}$ is defined by $J(\psi) = E\{|\psi(Z) - Y|\}$. The theorem stated below is based on the idea of (Lugosi and Zeger, 1995) which states the regularity conditions for universal consistency of the one hidden layered ANN model.

Theorem 1. *Let us take an ANN model with one hidden layer (also considering bounded output weights) having k hidden neurons and* σ *be a logistic activation function. Let* $F_{n,k}$ *be the class of neural networks defined as*

$$F_{n,k} = \left\{ \sum_{i=1}^{k} c_i\, \sigma\left(a_i^T z + b_i\right) + c_0 \,:\, k \in N,\ \ a_i \in R^{d_m},\ \ b_i, c_i \in R,\ \ \sum_{i=0}^{k} |c_i| \leq \beta_n \right\}.$$

Also, let ψ_n *be the function that minimizes the empirical* L_1 -risk *over* $\psi_n \in F_{n,k}$. *It can be proven that if k and* β_n *satisfy* $k \rightarrow \infty, \beta_n \rightarrow \infty, \frac{k{\beta_n}^2 \log(k\beta_n)}{n} \rightarrow 0$, *then the classification rule*

$$g_n(z) = \begin{cases} 0, & if \quad \psi_n(z) \leq 1/2 \\ 1, & \quad otherwise \end{cases}$$

is said to be universally consistent.

Equivalently, we write $J(\psi_n) - J^* \rightarrow 0$ in probability, where $J(\psi_n) = E\{|\psi_n(Z) - Y||\xi_n\}$ and $J^* = \inf_{\psi \in F_{n,k}} J(\psi_n)$. Write

$$J(\psi_n) - J^* = \left(J(\psi_n) - \inf_{\psi \in F_{n,k}} J(\psi) \right) + \left(\mathop{inf}_{\psi \in F_{n,k}} J(\psi) - J^* \right)$$

where, $\left(J(\psi_n) - \inf_{\psi \in F_{n,k}} J(\psi) \right)$ is called estimation error and $\left(\mathop{inf}_{\psi \in F_{n,k}} J(\psi) - J^* \right)$ is called approximation error.

To find the optimal choice of k of the proposed IEC model it is necessary to obtain the upper bounds on the rate of convergence (viz. how fast $J(\psi_n)$ approaches to zero) (Györfi et al., 2006). Though in case of the rate of convergence of estimation error, we obtain a distribution-free upper bound (Faragó and Lugosi, 1993). It is enough to find upper bounds of the estimation and approximation errors for finding k. The upper bound of approximation error was investigated by (Barron, 1993).

Proposition 1. *For a fixed* d_m, *let* $\psi_n \in F_{n,k}$. *If the proposed model satisfies the regularity conditions for the universal consistency as stated in Theorem 1, then the optimal choice of k is* $O\left(\sqrt{\frac{n}{d_m logn}}\right)$.

Proof. The upper bound of approximation error is found by (Barron, 1993) is $O\left(\frac{1}{\sqrt{k}}\right)$. The approximation error goes to zero as the number of neurons goes to infinity for universal consistency of the model, but for practical implementation the number of neurons is often kept as fixed. Using lemma 3 of (Faragó and Lugosi, 1993), we get the estimation error to be $O\left(\sqrt{\frac{kd_m \log(n)}{n}}\right)$.

Bringing them together, we get $J(\psi_n) - J^* = O\left(\sqrt{\frac{kd_m \log(n)}{n}} + \frac{1}{\sqrt{k}}\right)$.

For the optimal value of *k*, the problem reduces to equating $\sqrt{\frac{kd_m \log(n)}{n}}$ with $\frac{1}{\sqrt{k}}$, which gives $k = O\left(\sqrt{\frac{n}{d_m logn}}\right)$.

Remark. The optimal choice of hidden nodes in the universally consistent IEC model is found as $O\left(\sqrt{\frac{n}{d_m logn}}\right)$. But for practical application for small or medium-sized datasets, the recommendation of the paper is to use the number of hidden nodes in the hidden layer as $\sqrt{\frac{n}{d_m logn}}$ for achieving 'good' performance of the proposed model. The practical usefulness and competitiveness of the proposed classifier in solving a real life imbalanced business school data problem are shown in the next Section.

4. Application to Indian Business School Data

In this section, we first describe the business school data in brief and also discuss different evaluation measures that are used in this study. Subsequently, we are going to report the results of the experimentation and a comparison study on the proposed IEC model and other state-of-the-art classifiers.

4.1 Description of Dataset

The data was provided by a private business school that receives huge number of applications for the MBA course from across the country and admits a pre-specified number of students every year. This dataset comprises several parameters of last 5 years passed out students' profile along with their placement status. The dataset has 17 explanatory variables out of which 7 categorical variables and 10 continuous variables which represent the parameters of the students and one response variable, namely placement which indicates whether the students got placed or not (Chakraborty et al., 2018). In order to measure the level of imbalance of the data, we calculate the coefficient of variation (CV). Usually, dataset with a CV more than equal to 0.30 − a class ratio of 2:1 on a binary dataset is taken as imbalanced data. In the business school dataset, CV turns out to be 0.50. We also applied 5×2 cross-validation while evaluating various classifiers on the dataset, we have used 70% of the total available data as training and rest 30% of the data as testing. The experiments are repeated five times and the average results are reported in the paper. Table 1 provides an overview of the Indian business school dataset.

Table 1. Sample Indian business school dataset

ID	Gender	SSC Percentage	HSC Percentage	Degree Percentage	E-Test Percentage	SSC Board	HSC Board	Placement
1	Male	68.4	85.6	72	70	ICSE	ISC	Yes
2	Male	59	62	50	79	CBSE	CBSE	Yes
3	Male	65.9	86	72	66	Others	Others	Yes
4	Female	56	78	62.4	55.8	ICSE	ISC	Yes
5	Female	64	68	61	24.3	Others	Others	No
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

4.2 Performance Metrics

The performance evaluation metrics to be used in the experimental analysis are based on the confusion matrix. Higher the value of performance metrics, the better the classifier is. The expressions for different performance measures as follows:

$$\text{F-measure} = 2.\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}\,; \quad \text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}\,;$$

$$\text{G-mean} = \sqrt{\text{Recall}\times\text{Specificity}}\,; \ \text{AUC} = \frac{\text{Recall}+\text{Specificity}}{2}\,;$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}\,; \ \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}\,; \ \text{Specificity} = \frac{\text{TN}}{\text{FP}+\text{TN}}\,;$$

where, TP (True Positive): correct positive prediction; FP (False Positive): incorrect positive prediction; TN (True Negative): correct negative prediction; FN (False Negative): incorrect negative prediction.

4.3 Analysis of Results

Our aim is to select the optimal set of features and the corresponding model for the selection of the right set of students for the MBA program of a business school and subsequently will be placed as well. We compare our proposed imbalanced ensemble classifier (IEC) with other similar types of "imbalanced data-oriented" classifiers. Different performance metrics are computed to conclude from the experimental results. All the methods were implemented in the R Statistical package on a PC with 2.1 GHz processor and 8 GB memory.

We started the experimentation with HDDT algorithm by using R Package '*CORElearn*' for learning from imbalanced business school dataset. HDDT achieved around 93% accuracy while CT achieved around 83% accuracy. This indicates that "imbalanced data-oriented" classifiers perform better than the traditional supervised classifiers designed for general purposes. Further, we implemented HDRF, CCPDT which are among other imbalanced data-oriented algorithms. Finally, we applied our proposed imbalanced ensemble classifier which is a two-step methodology. In the first stage, we select important features using HDDT and record its classification outputs. Below are the important features we obtained for business school dataset by applying HDDT: SSC Percentage, HSC Percentage, Entrance Test Percentile, Degree Percentage, and Work Experience. In the next step, we design a neural network with the above mentioned important features along with HDDT output as an additional feature vector. The number of hidden neurons in the hidden layer of the model is chosen based on the recommendation of the proposed model (see Remark in Section 3). Min-max method is applied for scaling the dataset in an interval of [0, 1]. ANN training was done using '*neuralnet*' implementation in R. We reported the performance of different classifiers in terms of various performance metrics in Table 2. Table 2 shows that our proposed methodology achieved an accuracy of 96% for prediction in the Indian business school dataset.

Table 2. Quantitative measure of performance for different classifiers

Classifiers	**AUC**	**F-measure**	**G-mean**	**Accuracy**
CT	0.810	0.822	0.815	0.833
ANN	0.768	0.781	0.758	0.771
HDDT	0.933	0.936	0.925	0.931
HDRF	0.939	0.941	0.932	0.938
CCPDT	0.912	0.918	0.902	0.915
IEC	**0.964**	**0.969**	**0.951**	**0.960**

5. Conclusions

This paper proposed an imbalanced ensemble classifier (IEC) which took into account data imbalance and used it for feature selection cum imbalance classification problem. Through experimental evaluation, we have shown our proposed methodology performed well compared to the other state-of-the-art models. It is also important to note that "imbalanced data-oriented" algorithms perform well on the original imbalanced datasets. If we would like to work with the original data without taking recourse to sampling, our proposed methodology will be quite handy. IEC has the desired statistical properties like universal consistency, less tuning parameters and achieves higher accuracy than HDDT and ANN model. We thereby conclude that for the imbalanced business school dataset it is sufficient to use IEC model without taking recourse to sampling or any other imbalanced data-oriented single classifiers. Due to the robustness of the proposed IEC algorithm, it can also be useful in other imbalanced classification problems as well. Future work of the study will be to see the changes of the model in the presence of dataset shift in imbalanced scenario.

Conflict of Interest

The author declares that there is no conflict of interest for this publication.

Acknowledgement

The author would like to express his sincere thanks to the referees and editor for their valuable suggestions towards the improvement in the quality of the paper.

References

Barron, A.R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, *39*(3), 930-945.

Chakraborty, T., Chakraborty, A.K., & Murthy, C.A. (2019). A nonparametric ensemble binary classifier and its statistical properties. *Statistics & Probability Letters*, *149*, 16-23.

Chakraborty, T., Chattopadhyay, S., & Chakraborty, A.K. (2018). A novel hybridization of classification trees and artificial neural networks for selection of students in a business school. *Opsearch*, *55*(2), 434-446.

Cieslak, D.A., Hoens, T.R., Chawla, N.V., & Kegelmeyer, W.P. (2012). Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, *24*(1), 136-158.

Devroye, L., Györfi, L., & Lugosi, G. (2013). *A probabilistic theory of pattern recognition* (Vol. 31). Springer Science & Business Media.

Faragó, A., & Lugosi, G. (1993). Strong universal consistency of neural network classifiers. *IEEE Transactions on Information Theory*, *39*(4), 1146-1151.

Györfi, L., Kohler, M., Krzyzak, A., & Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Liu, W., Chawla, S., Cieslak, D.A., & Chawla, N.V. (2010). A robust decision tree algorithm for imbalanced data sets. In *Proceedings of the 2010 SIAM International Conference on Data Mining* (pp. 766-777). Society for Industrial and Applied Mathematics.

Lugosi, G., & Zeger, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Transactions on information theory*, *41*(3), 677-687.

Rao, C.R. (1995). A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *Qüestiió: quaderns d'estadística i investigació operativa*, *19*(1-3), 23-63.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Su, C., Ju, S., Liu, Y., & Yu, Z. (2015). Improving random forest and rotation forest for highly imbalanced datasets. *Intelligent Data Analysis*, *19*(6), 1409-1432.

Wang, L., & Alexander, C.A. (2016). Machine learning in big data. *International Journal of Mathematical, Engineering and Management Sciences*, *1*(2), 52-61.