Feature Selection for Machine Learning in Big Data

K.Kalpana, G.Sunil Vijaya Kumar, K.Madhavi

Abstract: ***We are in the information age there by collecting very huge volume of data from diverse sources in structured, unstructured and semi structured form ranging to petabytes to exabytes of data. Data is an asset as valuable knowledge and information is hidden in such massive volumes of data. Data analytics is required to have a deeper insights and identify fine grained patterns so as to make accurate predictions enabling the improvement of decision making. Extracting knowledge from data is done by data analytics, Machine learning forms the core of it. The increase in the dimensionality of data both in terms of number of tuples and also in terms of number of features poses several challenges to the machine learning algorithms . Preprocessing of data is done as a prior step to machine learning, so feature selection is done as a preprocessing step to have the dimensionality reduction of the data and thereby removing the irrelevant features and improving the efficiency and accuracy of a machine learning algorithm. In this paper we are studying various feature selection mechanisms and analyze them whether they can be adopted to sentiment analysis of big data.***

Index Terms:*Big data, Machine learning, Dimensionality reduction, Feature selection.*

I. INTRODUCTION

With important technological tendencies in the records era big content cloth of facts is being amassed and it is evolving in nature. statistics is gathered at a faster price from numerous resources like net, sensor networks, social networking web web sites, healthcare systems and masses of more. The statistics as a result accrued is of different formats from based totally, semi-established to unstructured. So we are living in the facts deluge age and this is remodeling[2] the corporation. massive statistics is defined with the resource of many V's[1] specifically by using manner of volume, variety, pace and Veracity. large statistics is big in volume, excessive speed and excessive in variety and it is mentioned the virtual information that is ever growing and difficult to control and examine the usage of conventional data analytics. system learning(ML)[4][5] performs a crucial role in the assessment of huge statistics. Many packages want effective analysis of records to find correlations, spot enterprise dispositions, are watching for disorder unfold, combat crime ,locating the fantastic of studies and plenty of greater. because of this the one of a kind packages of massive data embody enterprise system management, recommendation structures, sentiment assessment, health care analytics, smart towns, agriculture, energy grid control and many others. consequently massive information could make many changes in technology and generation and moreover in masses of factors of society.

One of the critical problem that is faced for analyzing Big data[3] is the presence of noise found with in it and as the data dimensionality increases the features contained in it also increases and finding the optimal features is some times an NP hard. Feature filter is applied to decrease the number of features there by doing the dimensionality reduction of data. Feature subset selection[11] and removal noisy and irrelevant features improves the efficiency and accuracy of an ML algorithm. Deep Learning[6] methods a part of ML techniques can be used in big data analytics and performance can be compared with the data mining algorithms.

II. BIG DATA CHALLENGES

The big information traits volume, range, pace and veracity poses severa stressful conditions to machine getting to know algorithms in records analytics[18]. amount is taken into consideration due to the truth the maximum amazing issue of massive information. top notch extent of information is accrued from numerous resources and amount of information is defined vertically through the extensive type of times and horizontally through manner of the huge style of attributes or the functions gift within the information set. due to the reality the volume of the facts increases the huge form of dimensions associated with it additionally growth. this situation is known as because of the truth the curse of dimensionality[13].

The performance and accuracy of the ML algorithm may be improved in big data environments by dimensionality reduction of the data through feature selection. Spurious correlations may also result in big data because of its enormous volume. ML algorithm also assumes that the whole training data has to be completely present in the memory for computation but it is not the same for the big data. And more over the performance of the ML algorithm also depends upon the architecture type on which it is running.

Another challenge in processing the big data is with its second characteristic that is variety where big data comes in different formats ranging from structured to unstructured and semi structured. Variety also refers to the different origins of information from wherein it's far accrued. various resources of facts include records from social network internet websites like twitter, facebook, blogs and lots of others. huge facts may also moreover includes lot of noise and outliers for this reason preprocessing of facts could be

Revised Manuscript Received on April 12, 2019.

K.Kalpana,Research scholar, Dept. of CSE, JNTUA (Email: kalpanagprec@gmail.com)

Dr.G.Sunil Vijaya Kumar,Professor, Dept. of CSE, G.Pulla Reddy Engineering College, Kurnool, A.P, India

Dr. K. Madhavi,Associate Professor, JNTUCE, Ananthapuramu, A.P, India.

very crucial in growing the overall usual performance of ML set of guidelines.

there can be a decentralized control in huge statistics computing environment and because the conventional ML algorithms assume that each one the statistics need to be present in memory, so the region of the statistics moreover becomes a undertaking. MapReduceparadigm[15] is used as a way to the above hassle, however MapReduce based totally completely answers face issues while jogging with the iterative algorithms.

Third distinctive attribute of big data is its velocity. Velocity of data is referred as the fastness at which data comes in. For example the data coming from a sensor and the messages posted on twitter etc. Real time fast processing of data is required for big data to take quick decisions. The ML algorithm must adopt to fast arrival of new data without retaining on the complete data set. So the ML algorithm must support incremental learning[8] as the data flow may not be continuous. The ML algorithms should consider the concept drifts[7] of big data.

The fourth aspect of Big data is its veracity which means the data is uncertain and imprecise as the data is collected from different un trusted sources. For example sentiment analysis done on data collected from social media is uncertain yet consists of valuable information .So in big data analytics ML algorithms has to deal with the ambivalent and imprecise data. To the above mentioned big data challenges there are several other challenges that need to be investigated and propose new techniques and methods to tackle them.

III. DIMENSIONALITY REDUCTION AND FEATURE SELECTION

Big data is incomplete and is susceptible to noise and consists of more number of dimensions. The accuracy of the ML algorithm is reduced because of high dimensionality and the presence of noise in big data. Hence preprocessing of big data has to be done before ML algorithm is applied on the data set. Dimensionality reduction[17] by feature selection is an important preprocessing step. The key benefit is many ML algorithms perform better if the dimensionality is reduced. Dimensionality reduction removes all irrelevant features and reduces noise in the data there by increasing the accuracy and efficiency of the classification algorithm and gets better quality clusters. Reduction of dimensionality leads to more understandable model and it allows the data to be more easily visualized. Finally the time and space complexities of the ML algorithm will be reduced.

Some of the approaches for dimensionality reduction is Principal Component Analysis(PCA)[20] is a linear algebra technique used for continuous data. Another linear algebra technique is Singular Value Decomposition(SVD)related to PCA is also used for dimensionality reduction.

Another method to reduce the dimensionality of data is by using the feature subset selection process[10][16].

By using this approach it may seem that we loosesome information but this is not the case when redundant and irrelevant featuresarepresentinthedata.Sometimesfeatureselection may be NP hard. Selecting the best features requiredomain knowledge and also a systematic approach because for *n* attributes there exists 2^n possible subsets. The search for finding the prominent subset of features may be very expensive as the *n* value and classes increases. Therefore some heuristic methods which implement greedy methodology areused.

There are three approaches to feature subset selection:

a) Embedded b) filter c) wrapper.

a) Embedded approach – Feature selection occurs as part of the ML algorithm. The algorithm decides which attributes to use and which attributes to ignore.

b) Filter approach - Features are selected systematically before the ML algorithm is applied.

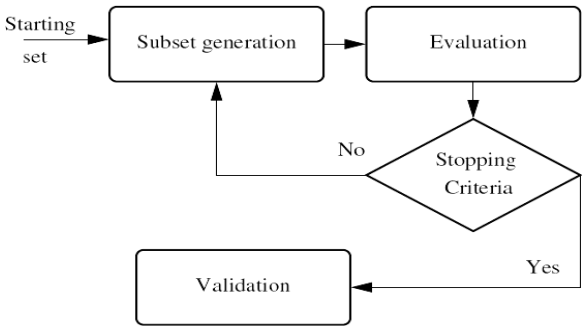
c) Wrapper approach[9]- These methods use the target ML algorithm as a black box to find the best features typically without enumerating all possible subsets.

form of function choice manner.

the overall shape of the function preference gadget includes the following steps.

Subset era, characteristic evaluation, stopping criterion, and validation [10].The function preference algorithms generate a first rate subset of abilties, evaluate them, and loops until the stopping criterion is happy [14]. in the end, the function subset diagnosed is tested thru the ML set of regulations on real information set.

feature subset choice is a are looking for over all possible subsets of features. the search approach must be computationally a lot less expensive and ought to find out most powerful features. There have to be evaluation manner to test the goodness of feature subset.



IV. RESULTS & DISCUSSION

One of the challenge that is faced by ML algorithms with big data is the presence of noise and high dimensionality. So as a preprocessing step to mining and to eliminate noise and select the best subset of attributes in big data we use Feature selection techniques[12].

Feature subset selection methods are both statistical and and entropy based. Various feature subset selection methods are

1. Chi-square
2. Analysis of Variance(ANOVA)
3. Information gain
4. Gain Ratio
5. One R

Chi-square: Pearson's Chi rectangular [21]is used for analyzing the unique facts. it's miles a statistical hypothesis

used as an example of a take a look at for independence between specific variables and a take a look at of goodness of healthy. The well really worth of a function is computed with the useful resource of using the chi rectangular with recognize to the splendor. The preliminary hypothesis H0 is idea that the two capabilities are unrelated and this is tested thru the chi squared technique

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

in this equation, 'O' is the found frequency, and 'E' is the predicted frequency. A immoderate X2 score charge does now not accept the null speculation of independence of the term and class. using Chi-squared to dataset and reducing the range of capabilities steadily permit us to take a look at how it may have an effect on the overall performance of ML set of guidelines like logistic regression. characteristic bargain improves the accuracy and standard overall performance of ML set of rules say for a logistic Regression. We can not say that Chi square is the nice method to growth the accuracy of a classifier.

assessment of variance (ANOVA): every different function choice technique this is used is analysis of variance(ANOVA)[22] a statistical method to determine if the way of or more impartial organizations is considerably one-of-a-kind. A score is assigned to each time duration based totally on the F-test. The pinnacle scored function phrases are considered because the first-rate preferred features and despatched to the ML set of policies. The device for F-test is given under.

$$F = \frac{Bg}{Wg} \quad (2)$$

Where Bg is the between group variability and wg is the within group variability

$$Bg = \frac{\sum_i ni(\bar{x}i - \bar{x})^2}{m-1}$$

$$Wg = \frac{\sum_{ij}(xij - \bar{x}i)2}{n - m}$$

In between-organization(Bg) variability ni is the total remember of observations of sophistication i, m is the full wide variety of instructions and $\bar{x}$ denotes the suggest of the information. In within-organization variability(Wg), xij denotes the jth remark in the ith class.

Entropy is used in information theory which is commonly used to separate the data and the samples are grouped into classes that they belong to. This is mainly used to maximize the groups purity. It is used in the calculation of the Information Gain(IG), Gain Ratio(GR), attribute ranking methods. System's unpredictability is measured by the entropy. The entropy of Y is

$$H(Y) = -\sum_{y \in Y} p(y) \log_2(p(y)) \quad (3)$$

in which p(y) is the opportunity of having the y rate whilst we're choosing one froma set. within the education facts set D if the values of Y are partitioned regular with the values of a different feature X, and the entropy rate of Y with apprehend to X rate is tons a great deal much less than the entropy of Y earlier than partitioning, then this represents a dating amongst capabilities Y and X.

statistics advantage(IG)

Given the entropy is a criterion of impurity in a education set D, we are able to outline a degree reflecting extra facts about Y furnished through X that would represent the amount thru which the entropy of Y decreases [23]. This degree is referred to as data gain. it is given with the resource of way of

$$IG = H(Y) - H(YX) = H(X) - H(XY) \quad (4)$$

IG is a symmetrical degree. The statistics gained about Y after looking X is equal to the statistics won approximately X after staring at Y. A drawback of IG is it's far biased in choice of abilities which have greater values even though they may be now not a bargain informative.

gain Ratio(GR)

to triumph over the downside of IG we've the benefit Ratio. it's miles a non-symmetrical measure[23]. gain Ratio is given with the aid of

$$GR = \frac{IG}{H(X)} \quad (5)$$

As equation (4) gives, normalization to IG is executed with the useful resource of dividing with the aid of the entropy of X, and vice versa while the variable Y have to be predicted. thru doing normalization the gain Ratio charge fall within the range[0,1]. If the price of GR is 1 it shows that the know-how of X truely predicts Y, and if the charge of GR is zero it technique that there may be no relation among Y and X. In assessment to IG, the GR works efficiently for variables with fewer values.

One-R

OneR is a primitive scheme proposed with the useful useful resource of Holte[24]. One rule is constructed for every function within the training statistics and it selects the rule of thumb of thumb of thumb of thumb having the smallest mistakes. Numerically valued skills are handled as non-stop and divides the kind of values into numerous disjoint periods. OneR handles missing values via treating "lacking" as a appropriate valid rate. It generates smooth rules through way of manner of the use of most effective one characteristic. OneR generates the recommendations which may be barely much much less accurate than the extraordinarily-current beauty algorithms however this produce hints which may be quite clean for humans to recognize and interpret.

Deep gaining knowledge of

Shallow mastering facts mining strategies like useful aid vector machines and preference wood aren't capable of extract the complex functions. In deep mastering models[19] hierarchical characteristic reading is used for extracting more than one layers of non linearabilties. The output of the remaining layer is used as capabilities for the classifiers or exquisite packages which mixes all the abilities for predictions.

V. CONCLUSION

As we are living in the digital age huge content of data is being collected at a fast rate. Identifying the knowledge hidden in big data is playing a major role in decision making. The traditional ML algorithms does not work efficiently for big data because of its characteristics. One of the main difficulty for processing big data is the

presence of

lot of noise and irrelevant features. Preprocessing has to be done to big data before machine learning is applied. So preprocessing step has to be done to big data by using a number of dimensionality reduction techniques that are suggested in literature. In this paper we studied to apply all the feature subset selection techniques on big data for doing sentiment analysis on social networking sites. The performance and accuracy of the ML algorithm can be increased by using the feature selection methods. As a future enhancement we can also apply the deep learning models to have the dimensionality reduction of big data.

REFERENCES

1. M.A.BeyerandD.Laney,''The importance of 'bigdata':A definition,'' Gartner Research, Stamford, CT, USA, Tech. Rep. G00235055, 2012
2. Mayer-Schönberger and K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think.
3. Boston, MA: Houghton Mifflin Harcourt, 2013
4. Chun-WeiTsai, Chin-Feng Lai, Han-ChiehChaoandAthanasiosV.Vasilakos "Big Data analytics: a survey Journal of Big Data (2015) 2:21
5. MachineLearning: What it is and why it matters.webLink:http://www.sas.com/en_us/insights/analytics / machine-learning.html
6. S. R. Sukumar, "Machine Learning in the Big Data Era: Are We There Yet?", ACM Knowledge Discovery and Data Mining: Workshop on Data Science for Social Good, Oak Ridge National Laboratory, pp. 1-5, December 2014.
7. M.M.Najafabadi,F.Villanustre,T.M.Khoshgoftaar,N.Seliya,R. Wald, and E. Muharemagic, ''Deep Learning Applications and Challenges in Big Data Analytics,'' J. Big Data, vol. 2, no. 1, p. 1, Feb. 2015.
8. P. B. Dongre and L. G. Malik, ''A review on real time data stream classification and adapting to various concept drift scenarios,'' in Proc. IEEE Int. Adv. Comput. Conf. (IACC), Feb. 2014, pp. 533–537.
9. X. Geng and K. Smith-Miles, ''Incremental learning,'' in Encyclopedia Biometrics. New York, NY, USA: Springer, 2009, pp. 731–735.
10. Kohavi, R., and John, G.H., "Wrappers for feature subset selection", *Artificial Intelligence*, 97 (1997) 273-324.
11. Dash, M., and Liu, H., "Feature selection methods for classifications", *Intelligent Data Analysis: An International Journal*, 1 (3) 1997.
12. http://www-east.elsevier.com/ida/free.htm.
13. Liu, H., and Motoda, H., *Feature Selection for Knowledge Discovery and Data Mining*, Kluwer Academic Publishers, 1998
14. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507–17
15. Grünauer A, Vincze M. Using dimension reduction to improve the classification of high-dimensional data. arXiv preprint arXiv:1505.06907. 2015.
16. Liu, H., and Setiono, R., "Chi2: Feature selection and discretization of numeric attributes",
17. Proc. IEEE 7th International Conference on Tools with Artificial Intelligence, 1995, 338-391.
18. Xing, E. P., Jordan, M. L., and Karp, R. M., "Feature selection for high-dimensional genomic microarray data", *Proceedings of the 18th International Conference on Machine Learning*, 2001, 601-608
19. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504–7.
20. A.Gandomi and M. Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, 35(2) (2015), pp.137-144
21. Ying He,F.RichardYu,NanZhao,VictorC.M.Leung,HongxiYin"Software-Defined Networks with Mobile Edge Computing and Caching for Smart Cities: A Big Data Deep Reinforcement Learning Approach"IEEE Communication Magazine(Volume: 55, Issue: 12, DECEMBER 2017)
22. Ding C, He X. K-means clustering via principal component analysis. In: Proceedings of the Twenty-first International Conference on Machine Learning, 2004, pp 1–9.
23. Pearson K. X. On the criterion that a given system of deviations from the probable in the case of a corsystem of variables is such that it can be reasonably supposed to have arisen from random sampling. Lond Edinburgh Dublin Philos Mag J Sci. 1900;50:157–75.
24. Fisher R. Dispersion on a sphere. Proc R SocLond. 1953;217:295–305.
25. Hall, M.A., and Smith, L.A., "Practical feature subset selection for machine learning",Proceedings of the 21st Australian Computer Science Conference, 1998, 181–191.
26. Holte, R.C., "Very simple classification rules perform well on most commonly used datasets",*Machine Learning*, 11 (1993) 63-91.