International Journal of Latest Trends in Engineering and Technology
Vol.(10)Issue(2), pp.381-386
DOI: http://dx.doi.org/10.21172/1.102.60
e-ISSN:2278-621X

SURVEY ON THE PRINCIPAL CHALLENGE OF TEXT MINING

Shweta Ganiger[1]

Abstract:Text mining is in a loosely systemized set of competing technologies that function as analytical with no clear dominance. It is the processing of unstructured text data. One of the main challenges is Dimensionality reduction, the dimensionality reduction is a process of reducing the number of random variables under consideration in a text document. It consists of two types feature selection and feature extraction. In this paper the singular value decomposition(SVD) and random forest technique of the dimensionality reduction are elaborated to know the how the dimensionality is reduced using these technique. The survey on dimensionality reduction problem proves that the random forest gives better accurate and performs well for documents. There are many other dimensionality reduction algorithms available to reduce the dimension of documents.

Keywords: Dimensionality Reduction, Feature selection, Feature extraction, SVD.

1. INTRODUCTION

Text mining or text analytics is the process of pulling out the interesting, non-trivial information and knowledge from unstructured data. Text mining and text analytics are broad umbrella terms describing a range of technologies for analyzing and processing semi structured and unstructured text data. The unifying theme behind each of these technologies is the need to “turn text into numbers” so powerful algorithms can be applied to large document databases. Converting text into a structured, numerical format and applying analytical algorithms require knowing how to both use and combine techniques for handling text, ranging from individual words to documents to entire document databases. Text mining can be divided into seven practice areas, based on the unique features of each area. However distinct, but these areas are highly interrelated; a typical text mining project will require techniques from many areas[6].

Text mining has many principal challenges such as Dimensionality reduction, noise mitigation, Intermediate form, Multilingual text refining. As the dimensionality reduction is became vast challenge of text mining[2].

Dimensionality reduction is the process of converting the large set of informative data having with vast dimensionality into less dimensionality without altering the meaning of informative data so that it can convey easily to readers or the reducing the number of random variables under consideration, via obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

Feature selection is also known as the variable selection or attributes selection. It is the method where the original variables are exacted using constructed methods. It consists some strategies they are filter for example information gain, wrapper approaches and embedded.

Feature reduction or feature extraction involves reducing the amount of resources required to describe a large set of data. The data transformation may be linear, as in principal component analysis(PCA), but many nonlinear dimensionality reduction techniques also exist.

As per the survey we found some of the dimensionality reduction methods:

Decision Tree: Decision tree can be used as solution to overcome multiple challenges like missing values, outliers and identifying significant variables, dimensionality reduction.

Principal Component Analysis (PCA): In this technique, variables are converted into a new set of variables, which are linear combination of original variables and these new set of variables are known as principle components. PCA uses singular value decomposition algorithm.

Random Forest: Similar to decision tree is Random Forest. Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees.

High Correlation: Dimensions exhibiting higher correlation can lower down the performance of model. Moreover, it is not good to have multiple variables of similar information or variation also known as “Multicollinearity”.

2. RELATED WORK

[1] School of Computer Science & Engineering, KLE Technological University, Hubbali, Karnataka, India

In the research, the text mining areas, its importance and main challenges of text mining are surveyed. The main challenge of text mining is dimensionality reduction; by using the various techniques feature selection and feature reduction are developed for dimensionality reduction of the documents. The comparisons between the techniques are discussed below.

Reddy, G. Suresh [1] has mentioned that dimensionality reduction has always been a main challenge in text mining, because it increases the complexity while mining a document with high dimension. The dimensionality reduction consist of scattered data, noise and space complexities. This problem is solved by feature selection and feature reduction methods. The author proposed some methods for this problem, the feature selection through singular value decomposition and information gain.

Method proposed by author in the paper novel membership function in is conventional Gaussian based. The similarity computed using this membership function for presence absence word pattern probabilities combination is same as that of the case where the word pattern probabilities are both present. After deriving the methods of dimensionality reduction novel member function, SVD the comparison is plotted in the graph. From the graph plotted for methods IG, considering 300 documents, the number of features retained after IG are 2836 , number of features after computing SVD are 223, and number of features initially present after pre-processing are 4220 which is very high. Of these, two approaches, it can be easily deduced that SVD is better compared to IG computation.

The author proposed a novel membership function for clustering document feature, this method is used to obtain optimal transformation matrix to achieve reduction in dimensionality of word documents. According to authors approach it proves that the proposed method is efficient and retains originality of words in document. Later the proposed method is compared with SVD and IG approaches, the proposed method is batter in comparison. The author concluded in future the work is to extend to various interrelated domains which requires dimensionality reduction, the domains such as network security, medical data analysis, and large dataset.

Zareapoor, et al.,[3] has taken a case study of phishing email detection, to conclude which dimension reduction methodis batter for classifying the text data. The email classification is the difficult task due to its huge sparse, noise and high dimensional feature. The two feature selection techniques chi-square and information gain, and two feature extraction methods Principal Component Analysis(PCA) and Latent Semantic Analysis(LSA) are used on the email dataset.

In feature extraction, the original feature space is converted to a more compact new space. All the original features are transformed into the new reduced space without deleting them but replacing the original features by a smaller representative set. That is when the number of feature in input data is too large to be processed then the input data will be transformed into a reduced representation set of features. PCA is a well-known technique that can reduce the dimensionality of data by transforming the original attribute space into smaller space. In the other word, the purpose of principle components analysis is to derive new variables that are combinations of the original variables and are uncorrelated. This is achieved by transforming the original variables Y = [y1, y2,...,yp] (where p is number of original variable) to a new set of variables, T = [t1, t2,..., tq] (where q is number of new variables), which are combinations of the original variables. Transformed attributes are framed by first, computing the mean (μ) of the dataset, then covariance matrix of the original attributes is calculated. And the second step is, extracting its eigenvectors. The eigenvectors (principal components) introduce as a linear transformation from the original attribute space to a new space in which attributes are uncorrelated. Eigenvectors can be sorted according to the amount of variation in the original data. The best n eigenvectors (those one with highest eigenvalues) are selected as new features while the rest are discarded. LSA method is a novel technique in text classification. Generally, LSA analyzes relationships between a term and concepts contained in an unstructured collection of text. It is called Latent Semantic Analysis, because of its ability to correlate semantically related terms that are latent in a text. LSA produces a set of concepts, which is smaller in size than the original set, related to documents and terms. It uses SVD (Singular Value Decomposing) to identify pat- tern between the terms & concepts contained in the text, and find the relationships between documents. The method commonly referred to as concept searches. It has ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts. LSA is mostly used for page retrieval systems and text clustering purposes. LSA overcomes two of the most problematic keyword queries: multiple words that have similar meanings and words that have more than one meaning. Information Gain is a feature selection technique that can decrease the size of features by computing the value of each attribute and rank the attributes. Then we simply decide a threshold in the metric and keep the attributes with a value over it. It just keeps those top ranking ones. Generally, Information Gain selects the features via scores. This technique can be simpler than the previous one. The basic idea is that we only have to compute the score for each feature that can reflects in discrimination between classes, then the features are sorted according to this score and then just keep those top ranking ones.

Data set is prepared by collecting a group of e-mails from the publicly available corpus of legitimate and phishing e-mails. Then the e-mails are labelled as legitimate and phishing correspondingly. The tokenization, stemming, missing value removal is performed in preprocessing. The experimental result shows the feature extraction methods performs batter classification on email dataset. Moreover the LSA method found to be best method in classification of the dataset and gives batter accuracy.

B.Azhagusundari, et al.,[4]proposes the attribute reduction the main process for knowledge acquisition. Here the author discussed about the algorithms on discernibility matrix and information gain to reduce the dimension. The efficiency and performance in classification of data is observed.

the selection method with information and discernibility presents good results in terms of number of feature selection and accuracy than applying methods individually. In future work the used algorithm will be tested on other dataset to explore more possibilities of methods of selecting optimal feature set. And to shows the dimensionality reduction and classification on large dataset.

Hideko KAWAKUBO, et al,.[10] they proposed the rapid feature selection method based on a empirical rule, the ranking of importance variable obtained from “Gini importance” and “mean decrease accuracy” differ, alike the members of top ranked variables in random forest are similar. The experimental ouput for evaluation time demonstrates the rapid feature selection is almost faster than the random forest technique while encountered with the high dimensional data.

3. ALGORITHMS WITH CASE STUDY

Feature selection using SVD and Feature extraction using Random Forest

SVD (Singular value decomposition):

SVD is not used to organize the data, its used to free the redundant data for dimensionality reduction. For example, if you have two variables, one is humidity index and another one is probability of rain, then their correlation is so high, that the second one does not contribute with any additional information useful for a classification or regression task. The eigenvalues in SVD help you determine what variables are most informative, and which ones you can do without.

Case study:Consider a set documents of CS and EC branch minors documents, these documents dimensionality is reduced by applying the SVD method.

Table.1.Set Document

Term	CO	IOT	DBMS	Mathematics	Electronics	CAM
Document						
CS-Minor1	4	4	4	0	0	0
CS-Minor2	3	3	3	0	0	0
CS-Minor3	1	1	1	0	0	0
EC-Minor1	0	0	0	2	2	2
EC-Minor2	0	0	0	3	3	3
EC-Minor3	0	0	0	1	1	1

SVD can be defined by: A[n x m] = U[n x r] ^[r x r] (V[m x r])T

A: n x m matrix (e.g., n documents, m terms)

U: n x r matrix (n documents, r concepts)

^: r x r diagonal matrix (strength of each ‘concept’) (r: rank of the matrix)

V: m x r matrix (m terms, r concepts)

solution

A = U ^ VT

U: document-to-document similarity matrix

V: term-to-document similarity matrix

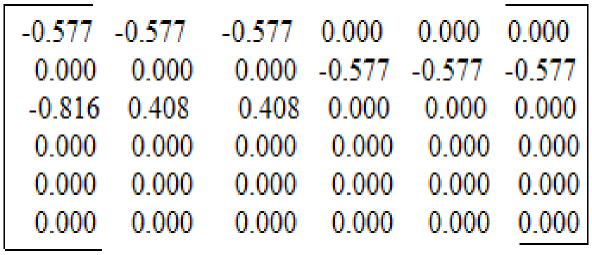
v12 = 0: data has 0 similarity with the 2nd concept

^: its diagonal elements: ‘strength’ of each concept

$$A=\begin{bmatrix} 4.000 & 4.000 & 4.000 & 0.000 & 0.000 & 0.000 \\ 3.000 & 3.000 & 3.000 & 0.000 & 0.000 & 0.000 \\ 1.000 & 1.000 & 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 2.000 & 2.000 & 2.000 \\ 0.000 & 0.000 & 0.000 & 3.000 & 3.000 & 3.000 \\ 0.000 & 0.000 & 0.000 & 1.000 & 1.000 & 1.000 \end{bmatrix}$$

$$A=\begin{bmatrix} -0.784 & 0.000 & 0.620 & 0.000 & 0.000 & 0.000 \\ -0.588 & 0.000 & -0.744 & 0.000 & -0.316 & 0.000 \\ -0.196 & 0.000 & -0.248 & 0.000 & 0.949 & 0.000 \\ 0.000 & -0.535 & 0.000 & 0.845 & 0.000 & 0.000 \\ 0.000 & -0.802 & 0.000 & -0.507 & 0.000 & -0.316 \\ 0.000 & -0.267 & 0.000 & -0.169 & 0.000 & 0.949 \end{bmatrix} \mathrm{X} \begin{bmatrix} 8.832 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 6.481 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{bmatrix} \mathrm{X} \begin{bmatrix} -0.577 & -0.577 & -0.577 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & -0.577 & -0.577 & -0.577 \\ -0.816 & 0.408 & 0.408 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & -0.816 & 0.408 & 0.408 \\ 0.000 & 0.707 & -0.707 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & -0.707 & 0.707 \end{bmatrix}$$

Now to reduce the dimentionaly of documents set the smallest singular value to zero



$$\begin{bmatrix} -0.577 & -0.577 & -0.577 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & -0.577 & -0.577 & -0.577 \\ -0.816 & 0.408 & 0.408 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{bmatrix}$$

The 6.481<8.832 so the document of EC-Minor1, EC-Minor2 and EC-Minor3 are reduced. This is how the dimensionality of the documents are reduced.

3.1 Random forest:

Random Forest is a versatile machine learning method capable of performing regression, classification and dimensional reduction methods, treats missing values, outlier values. Decision Tree Ensembles, also referred to as random forests, are useful for feature selection in addition to being effective classifiers.

Decision trees: Decision trees are the methods commonly used for data exploration.One of the type of decision tree is called CART(classification and regression tree).CART greedy, top-down binary, recursive partitioning, that divides feature space into sets of disjoint rectangular regions.Decision tree is encountered with over-fitting problem and ignorance of a variable in case of small sample size and large value. Whereas, random forests are a type of recursive partitioning method particularly well-suited to small sample size and large p-value problems.

WorkingTo categorize a new value based on given attributes, every tree given classification, its known as the tree "votes" for that class.

To classify a new object based on attributes, each tree gives a classification and we say the tree "votes" for that class. The forest chooses the classification having the most votes and in case of regression, it takes the average of outputs by different trees. The word 'Random' refers to mainly two process - 1. random observations to grow each tree and 2. Random variables selected for splitting at each node.

It works in the following manner. :

Assume number of cases in the training set is N. Then, sample of these N cases is taken at random but *with replacement*. This sample will be the training set for growing the tree.

If there are M input variables, a number m<M is specified such that at each node, m variables are selected at random out of the M. The best split on these m is used to split the node. The value of m is held constant while we grow the forest.

Each tree is grown to the largest extent possible and there is no pruning.

Predict new data by aggregating the predictions of the n trees (i.e., majority votes for classification, average for regression).

Case study:Assume Karnataka has a population of 6.0m, the Random forest algorithm can take up to 10k observation with only one variable to build each CART model. In this example 4 CART model are considered to build with different variables.

IncomeGroup :

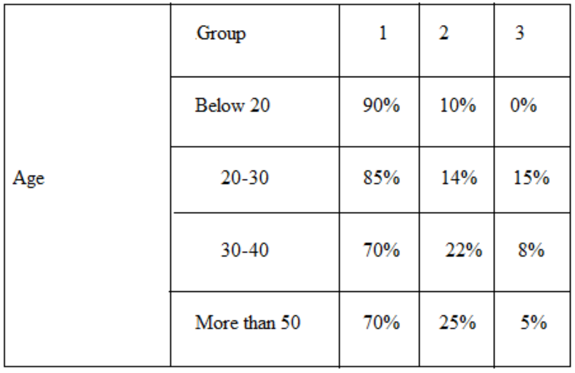
Income 1 : Below 20,000Rs

Income 2: 20,000 – 1,00,000Rs

Income 3: More than 1,00,000Rs

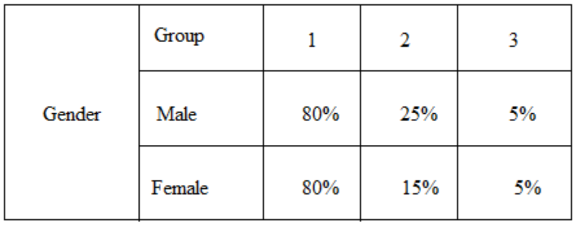
Following are the outputs of the 5 different CART model.

Table .2. CART 1 : Variable Age



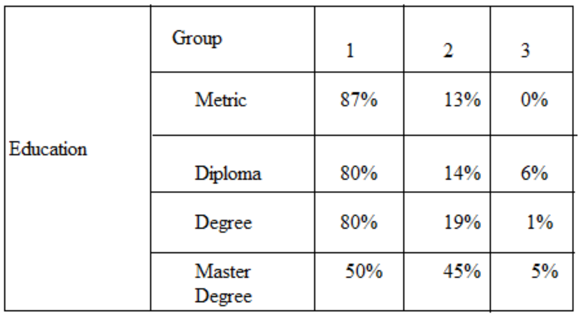
Age	Group	1	2	3
	Below 20	90%	10%	0%
	20-30	85%	14%	15%
	30-40	70%	22%	8%
	More than 50	70%	25%	5%

Table.3. CART 2 : Variable Gender



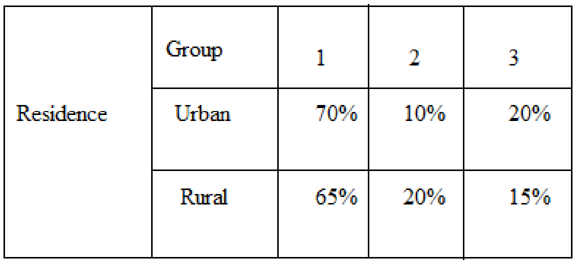
	Group	1	2	3
Gender	Male	80%	25%	5%
	Female	80%	15%	5%

Table.4. CART 3 : Variable Education



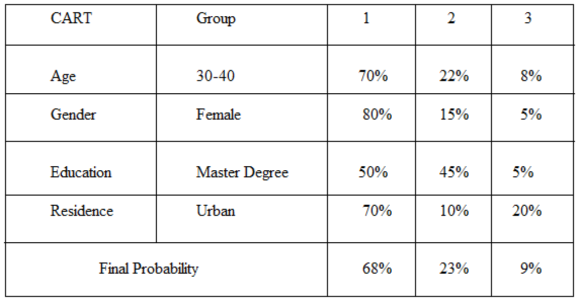
	Group	1	2	3
Education	Metric	87%	13%	0%
	Diploma	80%	14%	6%
	Degree	80%	19%	1%
	Master Degree	50%	45%	5%

Table.5. CART 4 : Variable Residence



	Group	1	2	3
Residence	Urban	70%	10%	20%
	Rural	65%	20%	15%

By using these 4 CART models, a single set of probability which belongs to the income class. This can be solved by "vote" method to calculate the final prediction. To calculate the final set of result considers following variables from each CART model:
Age : 30-40 years
Gender : Female
Highest Educational Qualification : Master Degree
Residence : Urban
For each of these CART model, following is the distribution across income:
Table.6.



CART	Group	1	2	3
Age	30-40	70%	22%	8%
Gender	Female	80%	15%	5%
Education	Master Degree	50%	45%	5%
Residence	Urban	70%	10%	20%
Final Probability		68%	23%	9%

The final probability is simply the average of the probability in the same income group in different CART models. As you can see from this analysis, that there is 68% chance of this individual falling in class 1, around 23% chance of the individual falling in class 2 and 9% in class 3. By these the case study the unwanted data which is not required is reduced and this shows how the Dimensionality reduction is carried out by Random Forest Algorithm.

4. CONCLUSION

In this paper we tried provide brief information of key challenge of text mining that is Dimensionality reduction, as the dimensionality reduction consists of many algorithms we came across our survey in this field. According to our survey we found two broad methods, feature selection and feature extraction. The feature selection method SVD and feature extraction method Random forest are deeply studied, from the survey it proves that the feature extraction gives good accurate results than feature selection. The Random forest algorithms outperforms in dimensionality reduction issue. It also used in classification, regression of the dataset.

5. REFERENCES

[1] [1].G.Suresh Reddy, " Dimensionality Reduction Approach for High Dimensional Text Documents", 2016 International Conference on Engineering & MIS(ICEMIS),DOI: 10.1109/ICEMIS.2016.7745364.

[2] [2].C. K. Chandrasekhar1, M. R. Srinivasan, B. Ramesh Babu," Bootstrapping in Text Mining Applications" ,International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611.

[3] [3].MasoumehZareapoor,Seeja K. R," Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection",I.J. Information Engineering and Electronic Business, 2015, 2, 60-65 Published Online March 2015 in MECS, DOI: 10.5815/ijieeb.2015.02.08

[4] [4]. B.Azhagusundari, Antony SelvadossThanamani," Feature Selection based on Information Gain", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-2, January 2013.

[5] [5]. Johannes Zenkert and MadjidFathi," Multidimensional Knowledge Representation of Text Analytics Results in Knowledge Bases", 978-1-4673-9985-2/16/$31.00 ©2016 IEEE 0541

[6] [6]. Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications "The Seven Practice Areas of Text Analytics"G. Miner, D. Delen, J. Elder, A. Fast, T. Hill, and R. Nisbet, Elsevier, January 2012.

[7] [7].Yanjun Li; CongnanLuo; Chung, S.M., "Text Clustering with Feature Selection by Using Statistical Data," in Knowledge and Data Engineering, IEEE Transactions on , vol.20, no.5, pp.641-652, May. 2008.

[8] [8]. Yung-Shen Lin; Jung-Yi Jiang; Shie-Jue Lee, "A Similarity Measure for Text Classification and Clustering,"IEEE Transactions on Knowledge and Data Engineering, vol.26, no.7, pp.1575-1590, July 2014.

[9] [9].SoniaLeach,"Singular value decomposition", Providence RI 02912, Draft version.

[10] [10]. Hideko KAWAKUBO, Hiroaki YOSHIDA," Rapid Feature Selection Based on Random Forests for High-Dimensional Data".

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

[12] [12].Seven Techniques for Dimensionality Reduction, Copyright © 2014 by KNIME.com