

A Comparative Study on Dealing with Sparsity in E-Commerce

¹Mohammed Ismail, ²C.H.S.L. Sowmya, ³T. Sai Sudheera and ⁴P Ravi Teja

¹Dept. of Computer Science and Engineering,

Koneru Lakshmaiah Education Foundation, AP, India.

mdismail@kluniversity.in

²Dept. of Computer Science and Engineering,

Koneru Lakshmaiah Education Foundation AP, India.

chlsowmya@gmail.com

³Dept. of Computer Science and Engineering,

Koneru Lakshmaiah Education Foundation, AP, India.

sudhi.tadiboina@gmail.com

⁴Dept. of Computer Science and Engineering,

Koneru Lakshmaiah Education Foundation, AP, India.

ravi9502088722@gmail.com

Abstract

Sparsity refers to the degree to which a dimension contains null values, or "NA"(Not Applicable). The NA values may either be missing or not applicable due to errors in network or any other issues like environment effects, malfunctioning of hardware etc. The impact of this is observed in many areas, especially in recommender systems like Google and e-commerce applications. It impacts the decision making of the customer. The existing mechanisms to overcome sparsity include usage of associative retrieval technology, collaborative filtering methods and few imputation methods. The problem with the existing mechanisms is that the absence of received value of a column is either a missing value or "NA" value. There

may be some cases where there is no need to fill the missing values rather leave them as "NA" value which is considered as a major problem and needs to be rectified. Also these existing mechanisms utilize large volumes of data which will continue increasing over time and will slow down system performance. The proposed system includes identifying null values and taking the decision whether to fill the null values or leaving it as NA value. We use probabilistic methods to determine whether the missing values are null values or NA values. Bayesian classification, one of the most popular methods is used here to take the decision regarding filling of missing values. A theorem called Expectation-Maximization (EM) theorem is used to fill the missing value because of its high accuracy in obtaining the results.

Key Words:Data sparsity, E-Commerce, expectation maximization theorem, bayesian classifier.

1. Introduction

In data science, data sparsity which is also known as data sparseness, data paucity etc. is the term used to describe the aspect of not having required or enough amount of information in Each data. Sparsity is of two types: Model sparsity and Data sparsity. [11]

Model sparsity doesn't create any impact as it means that there is a concise answer for the modeling effect whereas Data sparsity is usually treated as a problem which needs to be handled because the data that has gone missing might be very important data which may be useful for any further processing [13]. This paper deals with the comparative study of data sparsity problem whose impact is mainly observed in the applications over the internet like online sensor data, data present in websites specifically e-commerce applications.

The quick age of the Internet offered move up to the approach and method of electronic business (web based business), which has turned into a typical episode currently [12]. Internet-based commercial systems and different data bodies are the new business reality, as individuals and organizations reveal the ease of buying products and material services from different regions. The Internet business is characterized as electronic process through which people or associations can make distinctive exchanges, for example, purchase, and offer, exchange, or trade, merchandise, benefits and might be even data. As a result, e-commerce conclusively decreases the requirement of large investments or charge on substantial framework to establish an overall presence, which led to a innovation in a way business is organized across the globe.

In general, data in e-commerce consists of features, reviews, rating, and images about the product [17]. The problem of sparsity is reflected on the above parameters. Here the sparsity refers to the degree to which a dimension contains null values, or "NA"(Not Applicable). The NA values may either be missing or not applicable due to errors in network or any other issues like environmental issues, malfunctioning of hardware, or malfunctions during data or information exchange etc. One of the major problems include decision making ability of the customer, that is, to decide whether to buy the product or not, as the customer wouldn't be fully aware of the features [18]

To overcome this problem, we propose a solution which identifies the sparse data and decides whether to fill the data or leave it unchanged (Not Applicable). We use conditional probability to determine whether to fill the data or leave it unmodified. Based on the result obtained by the conditional probability, if the data needs to be filled, we use a theorem named Expectation-Maximization theorem to obtain accurate result.

2. Literature Survey

Information Filtering method in “Information Filtering in Sparse Online Systems: Recommendation via Semi-Local Diffusion” [1]

The authors of this research have proposed two personalized algorithms which deal with the sparse data. The sparsity in their problem has occurred while giving the recommendations to the customers. Here the recommendation data which is considered is Amazon and Book cross data. Normally we use a measure called ranking score to measure the accuracy of the recommendation. The lower the ranking score the higher will be the rate of accuracy. In this paper instead of considering ranking score as a measure, they have used a measure called recall because practically, users will only consider the top recommendations in a list. The coverage of diffusion increases non-linearly with the degree of the user which mainly increases the sparsity in the data set seriously. The proposed system of the semi-local diffusion increases coverage of the diffusion and decreases the sparsity by breaking the bond among the data items which are sparse.

Association Retrieval Method in “Solving the Sparsity Problem in Recommender Systems Using Association Retrieval” [2]

Collabrative filtering is a standout amongst the best approaches to diminish the sparsity issue while offering suggestions to the clients in E-Commerce. This approach gives recommendation to the current customer based on his past transactions and also by considering the feedback of the similar customers. But it is not that easy to find the common interests in between sparsity problem. Hence, in this method, the authors identified both direct similarity and indirect similarity between the customers, and computed a matrix called similarity matrix using a technique called association retrieval technology to understand the transitive associations that laid in the data based on the customers' feedback.

The informational index they considered in this paper was movie lens information. The principle behind this method is to construct a diagram or a system i.e a model of information that files the questions and terms and then recognize whether there is any transitive relationship among records and terms utilizing the considered chart demonstrate through which the nature of proposal will be enhanced and will likewise have great scope rate.

Decision Rule Method in “Decision-Rule Solutions for Data Mining with Missing Values” [3]

In this method the authors have used a decision rule approach but not in a traditional style. The methodology used includes formation of disjunctive normal rules (DNF). It starts with the Formation of the induced tree. Each tree node or class has an equal number of unweighted rules. The disjuncts are

performed and joined with the class having the missing values [14]. The disjuncts get matched or overlapped and the missing values' substitute will come into place. This method is a transparent approach for achieving the missing values. The main disadvantage is that the operations of pruning and optimization of nodes in later steps become very difficult.

Classifier method in “Improving Accuracy Rate of Imputation of Missing Data Using Classifier Methods” [4]

In this method, in order to fill the missing values in considered data set, the authors have used different imputation methods and checked their influence on performance of different classifiers when different subsequent imputed data is used. The different classifiers used in this paper are RIPPER, LSVM, SVM, SVMR, KNN and C4.5. The above mentioned classifiers are used to increase the rate of accuracy and they are well known imputation techniques [15]. After conducting experiments the results generated show us that among all the imputation methods used, MMSD method is better compared to all other imputation techniques. This method gives good performance when up to 20% of the data set goes missing.

K-medoids Method in “A K-medoids Algorithm Based Method to Alleviate the Data Sparsity in Collaborative Filtering” [5].

In this research, a strategy which utilizes K-medoids calculation in a method called collaborative filtering is proposed. Essentially this strategy contains three stages. Initial step incorporates bunching of assets utilizing K-medoids calculation. Second step incorporates buildup of client conduct information in view of the assets that have been grouped and last advance incorporates age of proposal list utilizing client based cooperative calculation and consolidated client particular information. The data set is considered from Internet source on platform of Internet education resources. This method provides significant development on both recall measure and precision of the considered data set. The future work of this paper includes concentrating on reducing the time complexity of the algorithm when a large data set is considered.

Imputation Methods in “Techniques to Deal with Missing Data” [6]

This method deals with the real-time data which means it contains many missing values in it. The implementation of this paper has three phases. First phase includes selection of features. Second phase includes filling missing values by making the use of various probabilistic methods and lastly the third phase includes correcting the values that have been filled in second phase. In this paper, imputation methods namely K-NN method and median & mean imputation method are used. The only one consideration is that, data set should consist of only numeric or nominal data. In results, the performance of both the algorithms used are compared using N-Fold cross validation methods using various different classifiers namely naïve-Bayes and so on. Both the methods are efficient and give maximum & same accuracy.

Decision Table Method in “An Algorithm: Filling the Missing Values in the Incomplete Decision Table” [7]

In this method, the authors proposed a new algorithm in order to fill the missing values in a decision table which is incomplete. This algorithm is the extension of another algorithm called ROUSTIDA algorithm. Most of the tables which contain information are incomplete because of errors. There are two approaches to handle these incomplete tables. Direct disposals broaden relative ideas of harsh arrangements of information hypothesis to inadequate table. Indirect transfer changes over incomplete table to finish filling. This paper looks at changed hypotheses of unpleasant sets and speaks about the contrasts between trait esteems in the table. In ROUSTIDA algorithm these differences are stored in matrix called discernibility matrix and is based on tolerance relation. If the classes which have limited tolerance misplace the indiscernibility classes in ROUSTIDA, then it can be improved and can be termed as a new, extended algorithm to ROUSTIDA algorithm. The new algorithm increases the consistency of the table.

Clustering Methods in “Cluster-Based Algorithms for Dealing with Missing Values” [8].

This research contains two phases. First one is to make review of existing mechanisms to deal with missing values and second one is to propose an algorithm which is clustered and based on mean & mode type of imputation method. In this paper the current systems to manage missing terms are gathered into two sets. Pre-replacement Methods: manage missing terms previously information mining preparing. Implanted Methods: handles the missing terms amid the way towards mining the information. The distinctive techniques for taking care of missing information are classified into these two sets. The primary thought of this cluster based calculations to fill missing terms is that as opposed to utilizing mean-mode on the aggregate dataset they utilized mean-mode in the subsets of entire dataset acquired by grouping. There are 3 variations for this mean-mode technique they are: Rank Cluster Based Mean-Mode algorithm (RCBMM), Natural Cluster Based Mean-Mode algorithm (NCBMM), K-Means Clustering based Mean-Mode algorithm (KMCMM)[16]. In all methods data set is first clustered and then implementation takes place. Experiment results prove that there is efficiency in execution time when clustered techniques are used.

Dynamic Attribute Selection Method in “Missing Value Estimation Based on Dynamic Attribute Selection” [9]

This method takes into consideration of missing values or data in the raw data. The authors proposed a new method to fill the missing data which selects the attributes one after the other in light of the shared data computed by smoothing the chosen terms. A decision tree is generated and it grows as values are added one by one. Firstly to use the information within the existing values class,

uncertainty is calculated, one by one and the previous values are aggregated to smoothen and use for the next missing value. The new attribute with maximum value is dynamically selected and used for later decisions. An advantage of this method is that it can be used where the missing values are high in ratio and correspondingly improves recognition of the missing values.

Bayes and Fuzzy Classifiers in “On Filling-in Missing Attribute Values for Bayes and fuzzy Classifiers” [10]

In this method they have compared two classifiers used for filling in the missing data which are naive Bayes and the fuzzy classifier. The decision rules used by the Bayes classifier and the type of fuzzy classifiers used are compared purely based on the scores and relative frequency in the training data set. The scores are calculated with a RVF classifier, a real valued classifier which is used for calculating scores based on real valued data. For multidimensional data the resulting score here is obtained by an aggregate operator. This paper has used the Bayes classifier for improving the low score values whereas the fuzzy set uses interpolation method to do the same task. They have looked at the qualities in light of how much the arrangement is steady inside the general structure of that specific classifiers. The result analysis show that the Bayes classifier have a good groundwork for probalistic conditions and good for the low score value boosting and the fuzzy set classifier has a fill in procedure which is supported by the fuzzy set framework.

3. Conclusion

Methods mentioned above give us the absolute idea of how to deal with the data containing missing values. There are several methods for identifying missing values and for filling them accordingly. It is understood that still there is scope of improvement in terms of performance of these methods i.e., taking the decision whether to fill the missing values or leaving them as “NA” values[20] . Although the most recent technique proposed is applied on the recommender systems where there is no scope for any data to remain as “NA” value. If we consider e-Commerce data which contains a list of products and their features or characteristics, there is a possibility of any value being “NA” value. The recent techniques did not handle this type of situation. And so we are proposing a technique which helps us in deciding whether to fill the missing value or making it “NA” value.[19] We are using a classifier called Naive bayes classifier to meet this requirement of taking appropriate decision. And after taking decision, in order to fill the missing value a theorem called Expectation-Maximization theorem is proposed which provides greater accuracy as it is iterative method. Thus the mentioned methodology provides the most accurate results for the considered data set.

References

- [1] Wei Zeng, An Zeng, Ming-Sheng Shang, Yi-Cheng Zhang, Information Filtering in Sparse Online Systems: Recommendation via Semi-Local Diffusion, PLOS ONE November 8(11) (2011).
- [2] YiBo Chen, ChanLe Wu, Ming Xie, Xiaojun Guo, Solving the Sparsity Problem in Recommender Systems Using Association Retrieval, Journal of Computers 6(9) (2011).
- [3] Weiss S., Indurkha N., Decision-rule solutions for data mining with missing values, Advances in Artificial Intelligence (2000), 1-10.
- [4] Thirukumaran S., Sumathi A., Improving accuracy rate of imputation of missing data using classifier methods, 10th International Conference on Intelligent Systems and Control (2016), 1-7.
- [5] Lin Ziqi, Ni Wancheng, Zhang Haidong, Zhao Meijing, Yang Yiping, A K-medoids Algorithm Based Method to Alleviate the Data Sparsity in Collaborative Filtering, Proceedings of the 34th Chinese Control Conference (2015).
- [6] Sessa J., Syed D., Techniques to deal with missing data, 5th International Conference on Electronic Devices, Systems and Applications (2016), 1-4.
- [7] Wu M.F., Wang T.L., An algorithm: Filling the missing values in the incomplete decision table, International Conference on Machine Learning and Cybernetics 1 (2008), 190-195.
- [8] Fujikawa Y., Ho T., Cluster-based algorithms for dealing with missing values, Advances in Knowledge Discovery and Data Mining (2002), 549-554.
- [9] Lee K.C., Park J.S., Kim Y.S., Byun Y.T., Missing value estimation based on dynamic attribute selection, Pacific-Asia Conference on Knowledge Discovery and Data Mining (2000), 134-137). Springer, Berlin, Heidelberg.
- [10] Ralescu A., Visa S., On filling-in missing attribute values for Bayes and fuzzy classifiers, Annual Meeting of the North American Fuzzy Information Processing Society (2008), 1-6.

- [11] Zeinab, Sharifi, A New Algorithm for Solving Data Sparsity problem Based-On Non Negative Matrix Factorization in Recommender Systems, 4th International Conference on Computer and knowledge engineering (2014).
- [12] Parveen R., Jaiswa A.K., Kant V., E-Learning Recommendation Systems–A Survey, International Journal of Engineering Research and Development 4(12) (2012), 10-12.
- [13] Reshma R., SAMbikesh G., Santhi Thilagam P., Alleviating Data Sparsity and Cold Start in Recommender Systems using Social Behaviour, Fifth International Conference on recent trends in Information Technology (2016).
- [14] Linden G., Smith B., York J., Amazon. com recommendations: Item-to-item collaborative filtering, IEEE Internet computing 7(1) (2003), 76-80.
- [15] Remya Menon, Shruthi S Nair, Srindhya K., Kaimal M.R., Sparsity-based Representation for Categorical Data, IEEE Recent Advances in Intelligent Computational Systems (2013).
- [16] Yan Hu, Qimin Peng, Xiaohui Hu, A Time-Aware and Data Sparsity Tolerant Approach for Web Service Recommendation, IEEE International Conference on Web Services (2014).
- [17] Li P., Zhang B., Weng Y., Rajagopal R., Autoregressive model for individual consumption data-Sparsity recovery and significance test, Power and Energy Society General Meeting (2016), 1-5.
- [18] Piccart B., Struyf J., Blockeel H., Alleviating the sparsity problem in collaborative filtering by using an adapted distance and a graph-based method, Proceedings of the 2010 SIAM International Conference on Data Mining (2010), 189-198.
- [19] Panahi A., Viberg M., Performance Analysis of Sparsity-Based Parameter Estimation, IEEE Transactions on Signal Processing 65(24) (2017), 6478-6488.
- [20] Li X., Girin L., Horaud R., Gannot S., Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25(10) (2017), 1997-2012.

