International Journal of Research in Engineering, Science and Management
Volume-3, Issue-2, February-2020
www.ijresm.com | ISSN (Online): 2581-5792

A Review on Data Reduction Techniques

Ansari Saad Aamir Javeed Akhtar[1], D. M. Kanade[2]
[1]*Student, Dept. of Computer Engineering, K. K. Wagh Institute of Engineering and Research, Nashik, India*
[2]*Professor, Dept. of Computer Engineering, K. K. Wagh Institute of Engineering and Research, Nashik, India*

***Abstract*: In data reduction unimportant data, noisy data, repetitive data and less important data is removed without affecting the original distributions of data. The data can be reduced in terms of instances (rows) and attributes (columns). In this paper various techniques related to data reduction are discussed with its application and limitation. Based on the study a new system requirement is proposed.**

***Keywords*: Data reduction, Dimension reduction, Feature selection, Granulation, Important labeling, KNN.**

1. Introduction

Due to heavy use of digitalized system, there is explosive growth in data volumes. A big data processing gained importance due to such heavy growth in data. Data collected in industries, organization, scientific domain, social sites etc. need to be processed and important data need to be mined. The efficiency of machine learning algorithm hampers while dealing with such big data. Big data also suffers from memory storage issue.

The solution for such big data problem is data reduction. In data reduction unimportant data, noisy data, repetitive data and less important data is removed without affecting the original distributions of data. This leads to less memory storage and low computational cost. The data reduction is a pre-processing step. It is applied before applying any machine learning algorithm.

The data can be reduced in terms of instances (rows) and attributes (columns). In instance data reduction, data instances i.e. records in a data are deleted whereas in attribute reduction attributes i.e. dimensions of data are deleted.

There are 2 main strategies for reduction: filter and wrapper. In filter, a selection metric is defined. The metric is defined on the basis of some clusters or marginal points. After data removal it checks the distance variation. After finding the subset of data the performance of reduced data is checked with the help of machine learning algorithms. Following figure shows the block diagram of filter method.

Fig. 1. Filter method

In wrapper technique, classification algorithm is used for data selection. This is an iterative method. Based on the inference drawn from the classification algorithm data is added or removed from the subset. Data which do not contribute in classification accuracy evaluation is deleted. Following figure shows the block diagram of wrapper method.

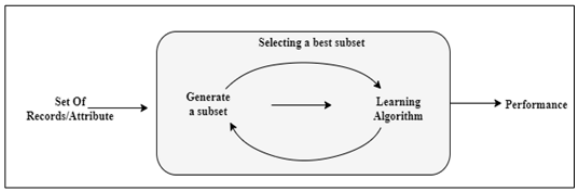

Fig. 2. Filter method

Lot of work has been done in data reduction domain. The attribute reduction also called as dimensionality reduction. The dimensionality reduction and instance reduction are two separate techniques. These techniques are studied independently and produce reduced data in terms of columns or rows. The combined approach can generate a better data representation.

The instance reduction method greatly reduces the size of data and improves the data storage efficiency. But it is difficult to manage tradeoff among classification accuracy, data reduction ratio and efficiency in reduction computation. Following section discuss the various data reduction strategies its working and limitations.

2. Literature survey

The data reduction methods are classified in two categories based on the technique used in it.

1) Wrapper Methods

The wrapper methods take the help of classifier for reduction purpose. Hart at. El. [2] proposes a wrapper based reduction algorithm. This algorithm is based on the nearest neighbor strategy. The system proposes a Condensed Nearest Neighbor (CNN) algorithm for instances selection based knn strategy. The system selects the instances and creates a subset of instances that correctly classify all instances in the original dataset. In this technique, the instance reduction rate is user defined and the system is unable to find the smallest representative subset. The system performance is dependent on a sequence of data.

GCNN [3] algorithm is proposed to improve efficiency of CNN algorithm. It finds the distance between nearest neighbor

instances and its nearest enemies. If the difference is higher than the threshold value then those instances can be deleted from the dataset. Wilson [4] proposes a technique named as edited nearest neighbor ENN. In this technique, instance is removed from the dataset by checking its class instances consistency with other dataset majority class instances using nearest neighbor algorithm.

Cuttlefish optimization algorithm [5] is used to reduce dataset instances. For efficiency improvement the principal component analysis algorithm is used. This is a feature extraction technique. It reduces the number of dimensions in the dataset. The reduced dimension set is used for instance selection process. The selected instances are deleted from original data with original attribute count. The feature set extraction improves the reduction process efficiency.

2) Filter Method

Filter method do not use any classification method for selection of candidates. The filter method uses some selection criteria.

Lumini and Nanni [6] proposes a clustering based data reduction technique. In this technique initially data is divided in number of clusters called as granules and the centroid of clusters are identified. The instances are selected from the cluster having less importance. The instances having less importance in cluster are deleted from the dataset.

J. A. Olvera-Lopez, et. al. [7] proposes a Prototype Selection Based Clustering (PSC) algorithm. This algorithm finds the representative instances from internal section of the class and all the boundary instances. It only deletes the internal class instances which are not labeled as important representative instances. This preserves the class covariance structure.

P. Hernandez-Leal et. al., [8] proposes a technique that ranks the instances on boundary section. The algorithm uses ENN algorithm to initially remove the noise in dataset. Then the instances are sorted based on the ranking score. The intra class section is retained. This paper tries to manage the tradeoff between classification time and accuracy in data reduction.

J. L. Carbonera and M. Abel [9] proposes a local density based instance selection method (LDIS). It evaluates the instances of each class separately. The dense area instances are preserved. The system focuses on complexity reduction.

Xiaoyan Sun, et. al. [1] proposes a data reduction technique based on the granulation process. This technique follows the combined approach of wrapper and filter method. The instances are initially granulated using k means algorithm. The instance importance is calculated on the basis of Hausdorff distance and crowding degree. To improve the efficiency of reduction process attribute mapping function is applied.

All the above techniques are instance reduction techniques. The dimension reduction techniques are studied independently in variety of papers. The dimension reduction techniques are generally used as a preprocessing task before applying any machine learning algorithm. An existing work, various filter [10] and wrapper [11] based techniques are used for dimension reduction. The dimension reduction and attribute reduction are studied independently in most of the papers.

3. Comparison among filter and wrapper methods

In most of the existing techniques filter and wrapper strategies are studied independently. In wrapper method, the data is divided in training and testing model and then importance of data is identified using some machine learning algorithm. In filter method there is no such data splitting required. This filter method uses some statistical formulae such as Euclidian distance, ranking, etc. for selecting subset of data whereas in wrapper methods cross validation is performed with the help of machine learning algorithms. The wrapper methods are iterative and hence require more time than the filter methods.

4. Conclusion

Data reduction is a preprocessing task that reduces the data size and keeps only important data. The data reduction is mainly categorized in 2 sections: Attribute reduction and Instance reduction. The efficiency of reduction algorithm is important. Various systems in literature are proposed to manage the tradeoff between efficiency of reduction and accuracy of reduction. The collective study and implementation of these techniques helps to reduce data size and improves the efficiency of machine learning algorithm.

References

[1] Sun Sun, Xiaoyan, Liu Lian, Geng Cong, and Yang Shaofeng, "Fast Data Reduction with Granulation based Instances Importance Labeling", IEEE Access, pp. 1-1.
[2] P. Hart, "The condensed nearest neighbor rule," IEEE Trans. Inf. Theory, vol. IT-14, no. 3, pp. 515-516, May 1968.
[3] C.-H. Chou, B.-H. Kuo, and F. Chang, "The generalized condensed nearest neighbor rule as a data reduction method," in Proc. Int. Conf. Pattern Recognit., Hong Kong, Aug. 2006, pp. 556-559.
[4] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," IEEE Trans. Syst., Man, Cybern., vol. SMC-2, no. 3, pp. 408-421, Jul. 1972.
[5] M. Suganthi and V. Karunakaran, "Instance selection and feature extraction using cuttlefish optimization algorithm and principal component analysis using decision tree," Cluster Comput., vol. 1, no. 2, pp. 1-13, Jan. 2018.
[6] A. Lumini and L. Nanni, "A clustering method for automatic biometric template selection," Pattern Recognit., vol. 39, no. 3, pp. 495-497, Mar. 2006.
[7] J. A. Olvera-López, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "Anewfast prototype selection method based on clustering," Pattern Anal. Appl., vol. 13, no. 2, pp. 131-141, 2010.
[8] P. Hernandez-Leal, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, and J. A. Olvera-Lopez, "Instance Rank based on borders for instance selection," Pattern Recognit., vol. 46, no. 1, pp. 365-375, Jan. 2013.
[9] J. L. Carbonera and M. Abel, "A density-based approach for instance selection," in Proc. IEEE Int. Conf. Tools Artif. Intell., Vietri sul Mare, Italy, Nov. 2015, pp. 768-774.
[10] Naoual El Aboudi, Laila Benhlima, "Review on wrapper feature selection approaches", in IEEE International Conference on Engineering & MIS (ICEMIS), Sept. 2016.
[11] K. Fathima Bibi, M. Nazreen Banu, "Feature subset selection based on Filter technique" in IEEE International Conference on Computing and Communications Technologies (ICCCT), Feb. 2015.