International Journal of Current Engineering and Technology
©2021 INPRESSCO®, All Rights Reserved

E-ISSN 2277 – 4106, P-ISSN 2347 – 5161
Available at http://inpressco.com/category/ijcet

Research Article

Attribute and Instance based Data Reduction Using Important Labeling

Ansari Saad Aamir Javeed Akhtar and Prof. D.M. Kanade

Department of Computer Engineering K.K. Wagh Institute of Engineering K.K. Wagh Institute of Engineering Education & Research ,Adgaon, Nashik,

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

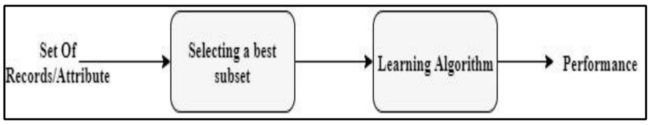
A bulk data is generated from various sources. The sources In Data reduction technique, the size of dataset is reduced by preserving important representative data. Lot of work has been done on data reduction techniques using machine learning. Data can be reduced in terms of attributes and instances. In existing approaches, the instance reduction and attribute reduction techniques are studied independently. The proposed system reduces the data in terms of attribute and instances. For attribute reduction, feature selection technique is used. Feature selection is a filter method that keeps only important attributes in dataset. For instance reduction, mapping based granulation and important instance labeling technique is used. Mapping technique reduces the multi dimensional data to the single dimensional data and then granules of one dimensional data are created using k means algorithm. Based on Hausdorff distance and data crowding degree, unimportant instances filtered from the dataset. The system will be tested on multiple UCI repository datasets. The efficiency of this system will be measured with the help of classification accuracy and execution time.

Keywords: *Data reduction, dimension reduction, feature selection, granulation, important labelling, knn*

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.

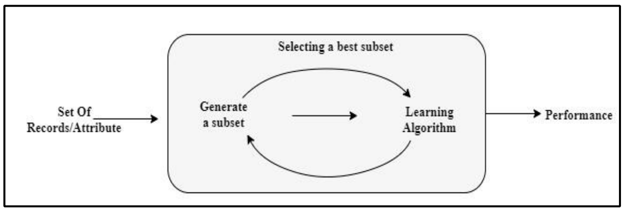
FIGURE1 : FILTER METHOD



.

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.

FIGURE 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. The proposed system is responsible for generating Reduced dataset by preserving the important data in dataset. The data is reduced in terms of attributes and instances. For attribute reduction, feature selection technique is used whereas for instance reduction, mapping based granulation and importance labeling is used. The system performance is measured in terms of classification accuracy and execution time.

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 dependant 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, at.el.[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 at, el [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, at. El. [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.

Problem Formulation

Let E1 be the real world entity. Re is set of records 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.

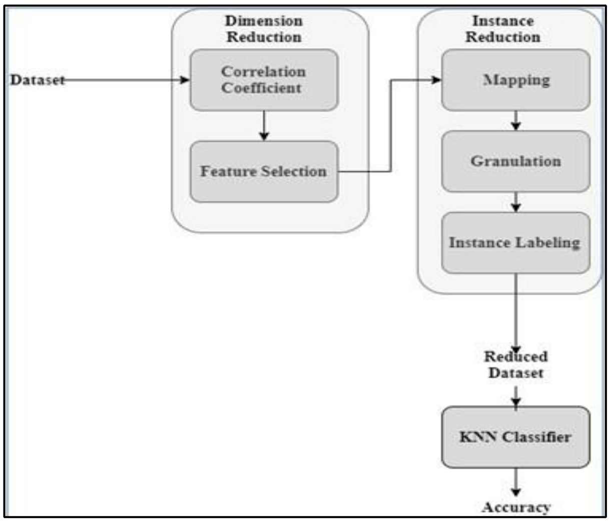
In most of the existing data reduction techniques filter and wrapper strategies are studied independently. In wrapper method, the data is divided in training and testing model and then importance of data is identified using some machine learning algorithm. In filter method there is no such data splitting 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.

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.

Proposed Methodology

Following Figure 3 shows the architecture of the system. The system takes dataset as an input. The reduction in terms of columns and row is performed. The reduced dataset is output to the system. The accuracy of reduced dataset is measured using knn classification technique.

FIGURE 3 : SYSTEM ARCHITECTURE



A. Preliminaries:

1. Correlation Coefficient :

P is a linear correlation measure between 2 attributes X and Y is given as:

$$P = \frac{\sum_i (Xi-Xm)(Yi-Ym)}{\sqrt{\sum (Xi-Xm)^2}\sqrt{(Yi-Ym)^2}_i}$$

Where, Xi Ɛ X and Yi Ɛ Y and Xm and Ym are the mean of attribute X and Y respectively.

The value of P lies between [-1, 1]. -1 or 1 value represents the complete correlation. The correlation value is 0 if two attributes are independent.

2. Mapping Function:

Let X {x1,x2,..xn} be the dataset with n number of instances and (xi1, xi2..xic, yi) be the number of attributes. The X dataset contains c+1 attributes, yi represent the class attribute.The c attributes in a data are mapped to one attribute using following equation:

$$T(xi)= \sqrt{\sum_{j=1}^{c} X_{ij}^2}$$

3. Distance Function

The distance between two instances after attribute mapping is calculated as:

Dpq = T(xp) – T(xq)

$$= \sqrt{\sum_{j=1}^{c} {x_{pj}}^2} - \sqrt{\sum_{j=1}^{c} xqj2}$$

Where T(xp) and T(xq) are the mapping values of instance xp and xq.

4. Hausdorff distance:

Let X be the dataset, X/xi is the dataset without xi instance. The Hausdorff Distance is calculated as:

H(X,X/xi) = max(h(X,X/xi), h(X/xi,Xi))

Where ,

h(X,X/xi) =||xp - xq|| , where Xp is the maximum value in X

and Xq is minimum value in X/xi

h(X/xi,Xi) = ||xp - xq|| , where Xp is the maximum value in X/xi and Xq is minimum value in X

5. Crowding degree:

The crowding degree is defined as:

$$Cd(xi) = \frac{1}{\sum_{x_j Ɛ X_\mu} ||xi-xj||}, \quad X\mu = \{xi \mid sig(xi)<=\mu\}$$

The large value of cd(xi) indicates that many instances are closer to the instance xi and xi can be removed. μ represents the threshold value.

B. System Working

The system is divided in two sections:

1. Attribute Reduction:

In attribute reduction, Ac numbers of attributes are removed from dataset based on the correlation coefficient value. The Attribute has low correlation value are removed from the dataset.

2. Instance reduction:

In instance reduction initially mapping function is applied and attribute count is reduced. The c numbers of attributes are reduced to one dimension using mapping function.

The mapping function improves the efficiency of reduction method. After mapping, granulation process is applied. For granulation, K-means clustering algorithm is applied and k different granules are generated. Rather than comparing the instance importance with whole dataset, importance of instance is calculated with the reference to granule in which it belongs.

The importance of instance is calculated using Hausdorff distance and crowding degree. If two instances in the dataset has similar values and very closer to each other then the Hausdorff distance between dataset X and dataset X/xi is smaller. It shows that if xi is removed then it has smaller influence on Xnew dataset. Instances having smaller importance than the given threshold value μ, can be deleted from the dataset. If two or more instances having same importance then crowding degree of those elements is calculated. Higher crowding degree represents lower importance. Numbers of instances are removed from the dataset based on reduction rate Ir given by the user. The accuracy of reduced dataset is calculated using KNN classification algorithm.

C. Algorithm:

Fast Data Reduction With Granulation Based Instances Importance Labeling and filter based attribute reduction
(FDR-GIILFAR)

Input: X: original dataset Ac: Attribute selection count Ir: Instance reduction rate **Output:** Xnew: reduced dataset**.**

Processing:

```
1.  CC: Calculate correlation coefficient of each attribute
2.  Select Ac attributes from dataset based on CC
3.  Apply mapping function to reduce dimensions space
4.  Generate K granules using Kmeans clustering algorithm
5.  Select random insistence xi
6.  Calculate Hausdorff distance (xi)
7.  If Hausdorff distance (xi)> μ
8.  Retain the instance
9.  else
10. Calculate the crowding degree of instances having same importance
11. delete the instances with larger crowding degree
12. update dataset as Xnew
13. If Ir count is not reached go to step 5
14. Return Xnew
```

Result and Discussions

A. Experimental Setup :

The System is implemented in java using jdk1.8. The system performance is evaluated on core i3 system with 4 gb ram.

B. Dataset :

UCI[12] benchmark data sets are used for system testing. Following table 1 describes the details of dataset in terms of number of attributes and number of instances.

Table 1 : Dataset Description

Sr. No.	Dataset	Number of attributes	Number of Instances
1.	Glass	10	214
2.	Liver	7	345
3.	Vehicle	18	846
4.	Iris	4	150
5.	Wine	13	178

C. Performance Measures:

1. **Accuracy:** Using KNN classifier and the accuracy of data reduction process is evaluated. The accuracy of original dataset and reduced dataset is compared.

2. **Evaluation Time:** Execution Time for data reduction process and classification process is evaluated.

D. Implementation Status:

The system is partially implemented. The correlation of each attribute is calculated with the class attribute and attribute having less correlation coefficient is deleted. The multidimensional data is then mapped to one dimensional array. The clustering is applied on one dimensional data and k granules are generated. Following table shows the time required for mapping and granulation process.

Table 2: Time Evaluation

Sr. No.	Dataset	Mapping process time in milliseconds	Clustering time in milliseconds
1.	Glass	117	548
2.	Liver	168	985
3.	Vehicle	267	1025
4.	Iris	98	342
5.	Wine	134	674

Conclusions

In this research work, the system works on data reduction in terms of attributes and instances. For attribute reduction, feature selection technique is used. For feature selection correlation coefficient is used. For instance reduction, mapping based granulation and importance labeling is used. The system performance is measured in terms of classification accuracy and execution time. In future, data reduction system can be implemented hybrid dataset containing numeric as well as nominal attributes.

References

[1]. Sun, Xiaoyan & Liu, Lian & Geng, Cong & Yang, Shaofeng, "Fast Data Reduction with Granulation based Instances Importance Labeling", IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2889122.

[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 nearestneighbor rule as a data reduction method," in Proc. Int. Conf. Pattern Recognit., Hong Kong, Aug. 2006, pp. 556-559.

[4]. 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]. 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. MartínezTrinidad, "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, "InstanceRank 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

[11]. Engineering & MIS (ICEMIS),Sept. 2016

[12]. 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

[13]. Dua and E. K. Taniskidou, "UCI machine learning repository," School Inf. Comput. Sci., Univ. California, Irvine, Irvine, CA, USA,

[14]. 2017. [Online]. Available: http://archive.ics.uci.edu/ml