Result Analysis for Instance and Feature Selection in Big Data Environment

Satish S. Banait, (Research Scholar) Department of Computer Engineering, KKWIEER, Nashik, India, ssbanait@kkwagh.edu.in

S. S. Sane, Department of Computer Engineering, KKWIEER, Nashik, Maharashtra, India,

Abstract: - Instance and feature selection has become an effective approach due to enormous data which is continuously being produced in the field of research. It is difficult to process such large datasets by many systems. Though the traditional techniques are useful for large datasets, the numbers when in hundreds, thousands or millions face scaling problems. The proposed work focuses on, scalable instance and feature selection in big data environment. Locality-sensitive hashing instance selection (LSH-IS) is a two pass method used to find similar instances along with Pearson correlation coefficient for feature selection. Hash function family is used which is a general method of reducing the size of a set; this is achieved by reindexing the elements into buckets. This process find similar instance in same bucket, hence instance can be reduced. The work aims at improving the performance of locality sensitive hashing by storing additional information of the instances and features assigned of each class in the bucket and also to improve accuracy of instance and feature selection algorithm.

Keywords — ***Big data, data reduction, feature selection, hashing, locality sensitive hashing, instance selection***

I. Introduction

Most of the data mining algorithms are applicable to small data sets with few thousands to lakhs of records. This degrades the efficiency of data being used for further processing. Presently, millions of records are the most scenarios; hence a new term emerged called as Big Data. Database sizes have grown considerably large in the recent years. Large sizes offer high challenges, which restricts machine learning algorithms to process such enormous volume of data and information. The significance of big data has nothing to do with amount of data you have, rather it deals with what to do with that data. Analysis of data from any resource can be done to find the answers for the facts that enable 1) minimum analysis of cost and reduction in time, 2) product growth, 3) efficient offerings, and 4) to make elegant decision. Merging of big data with high capacity analytics, accomplish task related to business such as:

- Find out defects, issues and the main reason of failure
- Efficient offerings at the point of sale based on the customers business practice
- To re-calculating total risk analysis within minutes
- Before the behaviour of an organization is affected detect faults

The quantity of data that's being produced and stored on a worldwide level is nearly unimaginable that keeps rising. It means that there is still even more likely to collect input insights from the business data and information, thus far, some amount of data is in fact used and analyzed. How does that suggest for analyst? What does this indicate for businesses? For businesses the unprocessed data and information that flows into organizations daily how they make good and efficient use of it?

In today's competitive complex business world various aspects of business are intermingled; to back up their decisions they need to rely on data. Large volume of data are collected and stored in databases, the requirement for efficient and effective analysis and utilization of the information contained in the data has been growing. Data sets that are accessible and available are progressively becoming huge in size, have difficulties in processing. Hence reduction techniques need to be applied. Different approaches are used by data reduction methods that includes instance selection, feature value discretization and feature selection. Data reduction is the procedure to minimize the amount of data that needs to be stored in a data storage background. It can reduce costs and increase storage efficiency. This work focuses on data reduction techniques such as instance and feature selection methods. The training set is reduced through instance selection which permits training stages of classifiers and also reducing runtimes in the classification. The process of selecting a subset of related features such as predictors and variables, that is for use in model construction is called feature selection. These methods are used for following reasons:

- To understand by researchers and users easily generalize the models
- Training time minimization

DOI : 10.35291/2454-9150.2022.0144

- Improve the model of generalization by dropping overfitting

Such data reduction techniques have emerged as substitute dominant meta-learning tool to accurately analyze the huge volume of data generated by modern applications. Due to such fast growth of such big data, solutions need to be studied in order to handle and extract value and knowledge from these data sets. Therefore, an analysis of the different types of data reduction techniques with big data sets may provide significant and useful conclusions.

II. Literature Survey

In recent years, data reduction analysis with improvement of algorithms has become the focus of a large amount of research effort. Very large number of data reduction algorithms has been developed for that purpose but none of the algorithm is suitable for all types of applications of data reduction analysis.

The nearest neighbour (NN) rule [1] [3] [2], in the training set it assigns a sample that is unclassified to the same class as the nearest of the N stored labelled samples. This rule is very easy, nevertheless powerful. The challenge to make an NN decision with an infinite number of samples is never worse than twice the Bayes risk [1]. To classify a test sample, large storage and computational requirements are enforced by NN method, as all the samples are labelled in the training set.

The condensed nearest neighbour (CNN) rule [4] is a variation of NN rule. It retains the similar and vital approach of the NN [1] rule, however it uses only a subset of the training set of samples. It is a two-stage iterative algorithm that is used for selecting a subset of a training set of samples which is used in a CNN decision rule. This subset correctly classifies all the samples that belong to the unique training set i.e. original for the NN decision rule, when it is used as a stored reference set. In CNN method boundary samples are occasionally retained rather internal samples are chosen randomly. In this way to add samples close to the decision tree, retention of interior samples is preserved in the condensed set.

In Prototype selection (PS) [5] the main approach is in comprehensive and large-scale image repositories reduce the number of training images. It is so to get better annotation performance and to make the most of reduction rate of sample sets. This PS algorithm is also named as DML-ENN that is Dissimilarity-based Multi-Label Edited Nearest Neighbor which reduces size of training set to overcome time complexity. When effective and useful training images by DML-ENN are found out, a well-known and fast classification method KELM known as Kernel Extreme Learning Machine [6] is used to enhance performance annotation. To predict label for unnoticed images this method is used to trained on reduced training sets.

In 1975 the authors planned a change to the meaning of a selective subset [7], for an enhanced estimate to decision borders. Although the condensed algorithm of Hart [3], is a subset of selective samples which can be thought of similar, but to apply a condition that is stronger than the consistency condition. Goal is easy, selected instances are found out, which is less responsive to the order of exploration of X and the random initialization of S in Harts [4] algorithm. Subset which is obtained is known as selective subset (SS), such that it satisfies the following conditions: 1) consistency, 2) the distance between any sample and its nearest selective neighbour within the same class is less than the distance from the sample to any sample of the other class, and 3) SS is as small as possible.

In LSs [8], is the origin of a supervised clustering algorithm. In instance selection (IS) method the results of this LS clustering were also used, included in a selective combination of IS methods. Most recently, in the framework of a meta-learning system, five different IS strategy [9] based on LSs were used. Few data-characterization measures based on LSs that conceive for systems which relate meta-learning to IS were used. For classification of new instance, LS gives a compressed explanation of the instance neighbourhood which is used to verify whether it is appropriate or not.

In online feature selection (OFS) [12], two different types of OFS tasks are addressed: 1) full inputs training, and 2) partial inputs training. In first task approach is assumed that the learner is able to access all the features of training instances. Here, for the correct prediction objective is to efficiently and clearly identify a fixed number of appropriate features. In other task approach is more difficult and challenging scenario is considered, where to identify the subset of relevant features for each training instance the learner is able to access a fixed small number of features. This problem is more attractable, because for each training instance it allows the learner to select fixed number of features that is to decide which subset of features to obtain.

Multi-criteria evaluation function [14] is used to characterize the importance of candidate features is proposed, by taking into thought not only the power in the boundary region and positive region but also their associated costs.

III. Architecture Design

A. System Architecture

The overall Block diagram of the proposed system is shown in the Figure 3.1

DOI : 10.35291/2454-9150.2022.0144

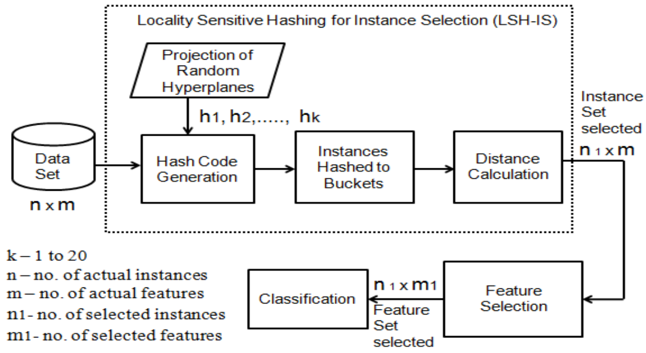
Locality Sensitive Hashing for Instance Selection (LSH-IS)
Projection of Random Hyperplanes
h1, h2,....., hk
Data Set
n x m
Hash Code Generation
Instances Hashed to Buckets
Distance Calculation
Instance Set selected
n 1 x m
Feature Selection
n 1 x m1
Feature Set selected
Classification
k – 1 to 20
n – no. of actual instances
m – no. of actual features
n1 - no. of selected instances
m1 - no. of selected features

Figure 3.1: System Architecture Diagram

The general architecture of the proposed system consists of several different stages as follows.

1. Dataset

Dataset is input to the system which is numerical. For the experiments, use datasets from the Keel repository that have at least 1000 instances.

2. Locality Sensitive Hashing for Instance Selection

The locality-sensitive hashing (LSH) is an efficient method for checking similarity between elements. It makes a particular use of hash functions that, unlike those used in other applications of hashing, seeks to allocate similar items to the same bucket with a high probability, and at the same time to greatly reduce the probability of assigning dissimilar items to the same bucket [16]. LSH use is common to increase the efficiency of nearest neighbors calculation.

3. Locality Sensitive Function

Given a set of objects S and a distance measure D, a family of hash functions H = h: S → U is said to be (d_1; d_2; p_1; p_2)-sensitive, if the following properties hold for all functions of h in the family H:

- For all x, y in S, if D(x,y) $\leq d_1$, then the probability that h(x) = h(y) is at least p_1.
- For all x, y in S, if D(x,y) > d_2 , then the probability that h(x) = h(y) is at most p_2.

The probability p_1 is associated with small distance d_1 and it is the lower bound on the probability of agreement for points at distance d_1 or less. The probability p_2 is associated with large distance d_2 and it is the upper bound on the probability of agreement for points at distance d_2 or more.

4. Hash Functions

The hash functions in the base family are obtained using the following equation,

$$h_{a,b}(x) = \left\lfloor \frac{a.x+b}{w} \right\rfloor \text{........ (1)}$$

Where, a is a random vector (Gaussian distribution with mean 0 and standard deviation 1), b is a random real value from the interval [0; w] and w is the width of each bucket in the hash table.

5. Pearson Correlation Coefficient

The Pearson correlation coefficient, frequently referred to as the Pearson r test is used for feature selection. It is a mathematical formula that evaluates the strength among variables and relationships. Coefficient value is used to find out how strong the relationship is between two variables.

$$r = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}}{\sqrt{\left(\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right)\left(\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}\right)}} \text{....... (2)}$$

Where,

n = number of pair of scores
ΣXY = sum of the products of paired scores
ΣX = sum of X scores
ΣY = sum of Y scores
ΣX^2 = sum of squared X scores
ΣY^2 = sum of squared Y scores

6. Classifier

Finally, selected set of instances and features will be given to the classifier for the classification performance. Three classifiers are used for this purpose namely 1NN, J48 and Adaptive Rule-Based (ARB) classifier.

7. Performance Evaluation

Classifiers performance will be evaluated base on the time taken by the classifier for classification. This is represented through graph.

B. *Algorithm*

For implementation of this system following algorithms have been used.

1. Locality Sensitive Hashing for instance selection (LSH-IS)
2. Pearson Correlation Coefficient for feature selection
3.

1. Locality Sensitive Hashing for instance selection (LSH-IS)

Input: A training set X = (x1, y1),....., (xn, yn), set G of hash function families

Output: The set of selected instances S ⊆ X

```
Processing:
(a) Initialize S = Φ
(b) foreach instance x Є X do
(c)   foreach function family g Є G do
          u ← bucket assigned to x by family g;
          Add x to u;
(d) foreach function family g Є G do
(e)   foreach bucket u of g do
(f)     foreach class y with some instance in u do
            $l_y$ all instances of class y in u;
            if $|l_y| > 1$ then
                Add to S one random instance of $l_y$;
(g) return S
```

DOI : 10.35291/2454-9150.2022.0144

2. Pearson Correlation Coefficient for feature selection

The Pearson correlation coefficient, frequently referred to as the Pearson r test is used for feature selection. It is a statistical formula that measures the strength between variables and relationships. Equation 2 represents this. In order to find out how strong the relationship is between two variables, a formula must produce what is referred to as the coefficient value. The range of coefficient value is in between -1.00 and 1.00. In negative range as one value increases, the other decreases. This means the relationship between the variables is negatively correlated. In positive range both values increase or decrease together. This means the relationship between the variables is positively correlated.

IV. Experimental Study

A. *Datasets*

For the initial experiment study, data sets are used from the Keel repository [17] which consists of at least 1000 instances. The major datasets are KDDCup99, CovType and Poker.

- KDDCup99 - We will use this dataset of year 1999 as an input training data for the projected work. This dataset contains 4,94,021 number of instances and consisting of 42 features.
- CovType - Dataset is Multivariate dataset of year 1998. This is used as an input training data for the projected work. This dataset contains 5, 81,012 number of instances and consisting of 54 features. Its attribute characteristics are categorical and integer.
- Poker - This dataset is Multivariate dataset of year 2007, which is used as an input training data for the projected work. This dataset contains 10,25,010 number of instances and consisting of 10 features. Its attribute characteristics are categorical and integer.

B. *Performance Evaluation*

For experiments we have used three main datasets that is KDDCup99, Cov- Type and Poker and some other datasets from the Keel repository [17] [18] that have at least 1000 instances. Table 4.1 summarizes the datasets: name and the accuracy given by three classifiers (using tenfold cross-validation): the nearest neighbor classifier with k=1, J48 and Adaptive Rule Based (ARB) Classifier and a classifier tree (weka implementation).

Methods	Classifiers	Datasets		
		KDDCup99	Cov-type	Poker
Instance Selection	1NN	99.0	94.2	51.0
	J48	99.0	95.09	70.0
	ARB	99.2	96.2	72.29
Feature Selection	1NN	99.5	94.2	51.4
	J48	99.5	95.05	70.2
	ARB	99.9	96.1	72.60
Proposed Method	1NN	99.95	95.6	52.2
	J48	99.95	96.0	72.1
	ARB	99.95	96.9	73.0

Table 4.1: Performance Comparison in terms of Accuracy (in percentage)

For different methods performance is evaluated in terms of accuracy. Tenfold cross validation was applied to the instance selection methods under study. The performance was as follows:

- accuracy achieved by 1NN, J48 and ARB classifiers trained with the selected subset.
- filtering time by instance selection.
- reduction achieved by instance selection methods (size of the selected subset).

C. *Execution Time*

The proposed system uses data reduction methods such as instance and feature selection so that effective and efficient data is achieved. Algorithm with linear complexity has been designed. The reduction rate can be improved by increasing or decreasing the no. of hash functions used. Due to sequential approach, computational cost is reduced. This instance and feature selection results in speed and low memory utilization hence are suitable for big data processing. Instance and Feature Selection is carried out on different datasets to reduce the timing and computational complexity caused due to large number of features.

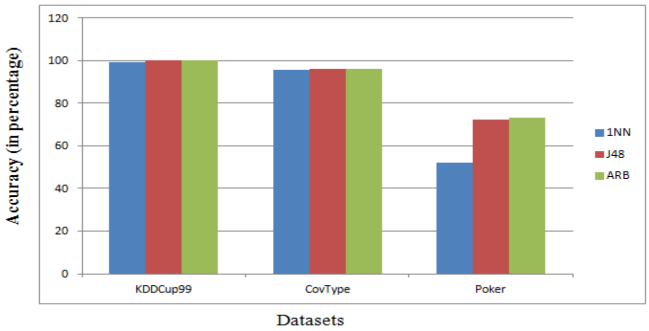

Figure 4.2 Accuracy of proposed method on Datasets using Classifiers

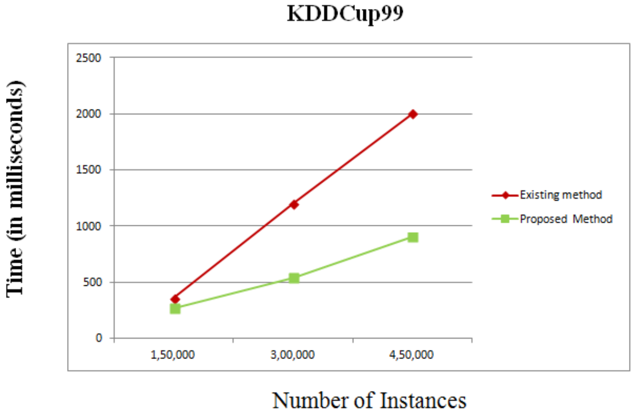

Figure 4.3: Execution time for KDDCup99 dataset for 1NN classifier

DOI : 10.35291/2454-9150.2022.0144

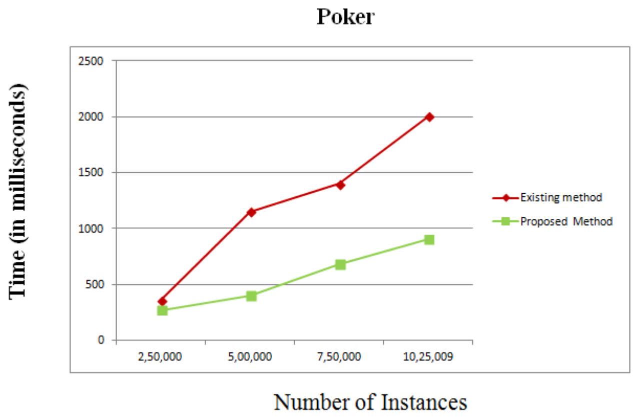

Figure 4.4: Execution time for KDDCup99 dataset for 1NN classifier

V. Conclusion

Data reduction of large spatial databases is very difficult task and also it requires high computational cost. The proposed system uses data reduction methods such as instance and feature selection so that effective and efficient data is achieved. It uses families of hash function to be generated for instance selection. Locality Sensitive Hashing Instance Selection (LSH-IS) needs two passes: in one pass it processes each instance of the data set and in second pass it processes the bucket of the families of hash functions. For feature selection Pearson Correlation Coefficient is used which is a statistical formula that measures the strength between variables and relationships. Hence instances and features that are similar are found out and proper set of instances and feature set is selected. Algorithm with linear complexity has been designed. The experimental results have shown that the strength of method is speed. The reduction rate can be improved by increasing or decreasing the no. of hash functions used. Due to sequential approach, computational cost is reduced. This instance and feature selection results in speed and low memory utilization hence are suitable for big data processing.

For experimental evaluation of our system, tests are conducted on various datasets with different classifiers. Experimental results evaluate the efficiency and effectiveness of our proposed technique.

References

[1] T. M. Cover and P. E. Hart, "Nearest neighbour pattern classification," IEEE Tram. Inform. Theoryv, ol. IT-13, pp. 21-27, Jan. 1967.

[2] Sujata Matale and S. S. Banait “A REVIEW ON INSTANCE AND

FEATURE SELECTION IN BIG DATA ENVIRONMENT” IJARIIE-ISSN(O)-2395-4396, Vol-3 Issue-2 2017.

[3] T. M. Cover, "Estimation by the nearest neighbour rule," IEEE Tram. Inform. Themy, vol. IT-14, pp. 50-55, May 1968.

[4] P. Hart, The condensed nearest neighbour rule (corresp.), Inf. Theor. IEEE Trans. 14 (3) (1968) 515-516.

[5] . Gates, The reduced nearest neighbor rule (corresp.), Inf. Theor. IEEE Trans. 18 (3) (1972) 431-433, doi: 10.1109/TIT.1972.1054809.

[6] Huang GB, Zhou H, Ding X, Zhang R (2012) Extreme learning machine for regression and multiclass classification. IEEE Trans Syst Man Cybern Part B Cybern

42(2):513-529.

[7] G. Ritter , H. Woodruff, S. Lowry, T. Isenhour, An algorithm for a selective nearest neighbor decision rule, IEEE Trans. Inf. Theor. 21 (6) (1975) 665-669.

[8] Y.Caises, A.Gonzalez, E.Leyva, R.Perez, Combining instance selection methods based on data characterization: an approach to increase their effectiveness, Inf.Sci.181(20)(2011)4780-4798.

[9] S. Garcia, J. Derrac, J. Cano, F. Herrera, Prototype selection for nearest neighbour classification: Taxonomy and empirical study, Pattern Anal. Mach. Intell. IEEE Trans. 34 (3) (2012) 417-435, doi: 10.1109/TPAMI.2011.142.

[10] E. Leyva, A. Gonzalez, R. Perez, Three new instance selection methods based on local sets: a comparative study with several approaches from a bi-objective perspective, Pattern Recognit. 48 (4) (2015) 1523-1537, doi: 10.1016/j.patcog. 2014.10.001.

[11] V. Bolon-Canedo, N. Sanchez-Marono, A. Alonso-Betanzos, Recent advances and emerging challenges of feature selection in the context of big data, Inf. Sci. 86(2015) 33-45

[12] Jialei Wang, Peilin Zhao, Steven C.H. Hoi, Online Feature Selection and Its Applications; IEEE Trans. Vol. 26, No. 3, March 2014

[13] P. Bermejo, L. Ossa, J.A. Gamez, J.M. Puerta, Fast wrapper feature subset selection in high-dimensional datasets by means of filter reranking, Knowl.- Based Syst. 25 (2012) 3544.

[14] C. Elkan, The foundations of cost-sensitive learning, in: Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI), 2001, pp.973978.

[15] E. Merelli, M. Pettini, M. Rasetti, Topology driven modeling: the is metaphor, Natural Comput. 14 (3) (2014) 421430, doi: 10.1007/s11047-014- 9436- 7.

[16] Farid, Mohammad, Bernard, Ann Nowe, An adaptive rule-based classifier for mining big biological data, Science Direct, 0957-4174 2016 Elsevier Ltd.

[17] Y. Hochberg, A sharper bonferroni procedure for multiple tests of significance, Biometrika 75 (4) (1988) 800-802, doi: 10.1093/biomet/75.4.800.

[18] J. Alcala-Fdez, L. Sanchez, S. Garcia, M. del Jesus, S. Ventura, J. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J. Fernandez, F. Herrera, Keel: a software tool to assess evolutionary algorithms for data mining problems, Soft Comput. 13 (3) (2009) 307-318, doi: 10.1007/s00500- 008- 0323- y.

DOI : 10.35291/2454-9150.2022.0144