A REVIEW ON INSTANCE AND FEATURE SELECTION IN BIG DATA ENVIRONMENT

S. M. Matale[1], S. S. Banait[2]

[1]*PG Student, Department of Computer Engineering, KKWIEER, Nashik, Maharashtra, India*
[2] *Assistance Prof, Department of Computer Engineering, KKWIEER, Nashik, Maharashtra, India.*

ABSTRACT

Instance and feature selection has become an effective approach due to the 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 F (LSH-IS-F) 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 and features in same bucket, hence instance/features can be reduced. The work aims at improving the performance of locality sensitive hashing by storing extra statistics of the instances and features that is assigned to each class in the bucket and also to improve accuracy of instance and feature selection algorithm by prototype generation.

Keyword: - *Big Data, data reduction, feature selection, hashing, instance selection*

1. INTRODUCTION:-

Most of the data mining algorithms are applicable to small data sets with few thousands to lacks 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

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
- 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.

In the below sections we are going to discuss about related work done for the proposed research area. We refer some existing research paper for completing this task. It is given as follow:

2. RELATED WORK

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 [l] [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 [l]. 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 [3] 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) [4] 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 [5] 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 [6], 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 [3] 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 [7], 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 [8] 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 Leyva et al [9], it focuses on instance selection for nearest neighbour classification. Without affecting the classification accuracy the goal is to decrease number of instances in the training set. Based on local sets, three instance selection methods are proposed, which follow complementary and various different methods, this is done in order to determine which instances to eliminate and which not. Among these one of method removes overlapped and noisy instances. Although it hardly decreases the database, but increases the accuracy more than any other method. Furthermore, it plays an vital task in the other two methods. In second method the approach is to select the centroid from clusters, and its main aim lies in reduction. Border instances are selected by the third method and achieve the greatest compromise among reduction accuracy. Though they have diverse reduction accuracy priorities, in the Pareto frontier of the reduction accuracy maximization problem non-dominated solutions is offered by all of them.

In online feature selection (OFS) [11], 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.

In this paper, authors Canedo, Marono and Betanzos [10] summarizes hot spots in feature selection research, intended at supporting scientific community to search for new opportunity and challenge that have newly arise. In variety of appliance it discusses the source and significance of feature selection and outline latest contributions, from DNA microarray analysis to face recognition. More years have witness formation of huge data sets and will constantly grow in size and number. The term big data situation present chance and confront to feature selection researchers, as a increasing and efficient need for scalable and efficient feature selection methods, given that the existing traditional methods are possible to prove insufficient.

FSNF [12] Nearest neighbours and Farthest neighbours aims at feature selection for clustering. This method achieves feature selection for clustering with no need to search the correct clustered information. Due to its robustness and well known, FSNF uses the mutual information criterion to determine salient features. The nearest and farthest neighbours help out to select the salient features; FSNF which uses the mutual information criterion to assess features at the same time considering these neighbours based on distinguish ability and redundancy toward a robust information assessor. In place of visiting the space of all feasible feature subsets, FSNS then determines a real-valued salience vector. Once a local-finest result which is salience vector is achieved, these salient features using learning algorithms are used to perform clustering.

In this, a multi-criteria evaluation function [13] 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. Using this, for selecting a feature subset of minimum cost that maintains the similar information as the complete feature set; forward greedy feature selection method is developed. Also implement the selection of candidate features in a dwindling object set to improve better efficiency of this algorithm.

3. SYSTEM ARCHITECTURE

The below figure 1 is the general system architecture of the proposed system.

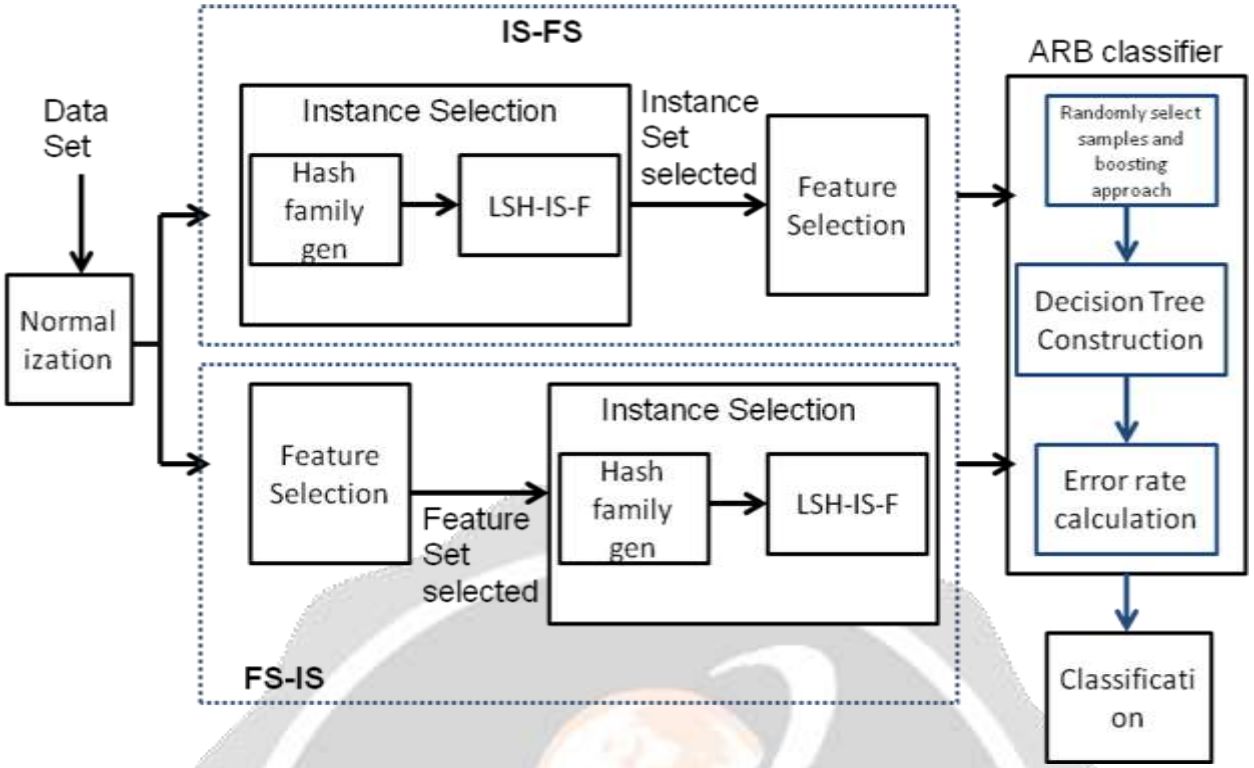

Figure 1: System Architecture

Data set is input to the system which is numerical. Initially normalization that is pre-processing of all input features is carried out to adjust values measured on different scales to a notionally common scale. Usually, large number of instances are accumulated and stored in the training set; all of instances are not needed for classifying. So to get adequate classification rates ignoring non useful cases, this process is called as instance selection. In this work Locality Sensitive Hashing Instance Selection F (LSH-IS-F) is used to reduce instances that are similar. This method works in two passes: in one pass each instance of the data set is processed and in second pass it processes the bucket of the families of hash functions.

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. 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 find out how strong the relationship is between two variables. There are two approaches to use instance and feature selection method for data reduction. Whether to use instance-feature selection (IS-FS) or feature-instance selection (FSIS) depends on the minimum value of error rate of method as compared to predefined threshold. To calculate error rate adaptive rule-based (ARB) classifier is used which works in three parts as: randomly selects samples from the input set, then proceed with decision tree (DT) construction and lastly calculate the error rate. The process whose error rate is low is used for further classification or any other business process, to evaluate the extent to which the instances selected by the algorithms were suitable for training other classifiers.

4. CONCLUSION

In this review paper, several existing techniques have studied and analysed in section II. Traditional methods of data reduction does not work effectively and efficiently for large amount of data. 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. For feature selection Pearson Correlation Coefficient is used which is a mathematical formula that evaluates the strength among variables and relationships. Hence instances and features that are similar are found out and are removed. The ARB classifier deals with class-imbalance problem, noisy instances and overfitting. According to our analysis, there is need of such system which has capability to overcome the shortage in existing system. From overall discussion and study analysis we analyzed that this instance and feature selection results in speed and low memory utilization hence are suitable for big data processing.

5. ACKNOWLEDGEMENT

It gives me an immense pleasure to express my sincere and heartiest gratitude towards our guide Prof. S. S. Banait, Department of Computer Engineering for his guidance, encouragement and moral support during the course of our work. He has proven to be an excellent mentor and teacher. This work is also the outcome of the blessings, guidance and support of our parents, family members and friends. Lastly our thanks to all who have contributed indirectly in completion of this work

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

[3] P. Hart, The condensed nearest neighbour rule (corresp.), Inf. Theor. IEEE Trans. 14 (3) (1968) 515516

[4] G. Gates, The reduced nearest neighbor rule (corresp.), Inf. Theor. IEEE Trans. 18 (3) (1972) 431433, doi: 10.1109/TIT.1972.1054809.

[5] 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):513529.

[6] G. Ritter , H. Woodruff, S. Lowry, T. Isenhour, An algorithm for a selective nearest neighbor decision rule, IEEE Trans. Inf. Theor. 21 (6) (1975) 665669.

[7] Y.Caises, A.Gonzlez, E.Leyva, R.Prez, Combining instance selection methods based on data characterization: an approach to increase their effectiveness, Inf.Sci.181(20)(2011)47804798.

[8] S. Garcia, J. Derrac, J. Cano, F. Herrera, Prototype selection for nearest neighbor classification: Taxonomy and empirical study, Pattern Anal. Mach. Intell. IEEE Trans. 34 (3) (2012) 417435, doi: 10.1109/TPAMI.2011.142.

[9] E. Leyva, A. Gonzlez, R. Prez, 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) 15231537, doi: 10.1016/j.patcog. 2014.10.001.

[10] 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) 3345

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

[12] 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.

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

[14] 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 .

[15] Farid, Mohammad, Bernard, Ann Nowe, An adaptive rule-based classifier for mining big biological data, ScienceDirect, 0957-4174/ 2016.

[16] Y. Hochberg, A sharper bonferroni procedure for multiple tests of significance, Biometrika 75 (4) (1988) 800802, doi: 10.1093/biomet/75.4.800.

[17] J. Alcal-Fdez, L. Snchez, S. Garcia, M. del Jesus, S. Ventura, J. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J. Fernndez, F. Herrera, Keel: a software tool to assess evolutionary algorithms for data mining problems, Soft Comput. 13 (3) (2009) 307318, doi: 10.1007/s00500- 008-0323- y.