

ENHANCING REDUNDANT HIGH DIMENSIONAL DATASETS WITH FEATURE SELECTION ALGORITHM

R. BHAVANA

M.Tech, Dept of CSE
IARE

brbhavana4@gmail.com

P. ANJIAH

Asst. Prof. Dept of CSE
IARE

anjaiah.pole@gmail.com

P.L. SRINIVASA MURTHY

Prof. Dept of CSE
IARE

plsrinivasamurthy@gmail.com

YERRAGUDIPADU SUBBA RAYUDU

Asst. Prof. Dept of CSE
IARE

subbu.iare@gmail.com

Abstract

Data retrieving in high dimensional information with few perceptions are ending up more typical, particularly in microarray information. Amid the most recent two decades, loads of effective arrangement models and FS (Feature Selection) algorithms have been proposed for higher forecast correctness. In any case, the result of a FS algorithm with considering expectation precision can be shaky among the varieties in the preparation set, particularly with high dimensional information. This paper suggests another assessment calculation Q-statistic that consolidates the solidness of the chose include subset notwithstanding the forecast precision. At that point, we propose the Booster of a FS algorithm that lifts the estimation of Q-statistic of the calculation connected. Observational investigations demonstrate that Booster helped in the estimation of the Q-statistic as well as the expectation exactness of the calculation connected unless the informational index is characteristically hard to anticipate with the given algorithm.

Key Words—Accuracy, Prediction algorithms, Redundancy, Q-statistic, FS, Booster

1. INTRODUCTION

The advent of various new application domains, such as bioinformatics and e-commerce, health care and education excetra, underscores the necessitate for scrutinizing high dimensional data. Thus mining high dimensional data is an compelling plight of exceptional pragmatic significance. Obviously, mining of data (once in a while called data Feature Selection[1][2] (FS) is applied to lessen the number of features (attributes) where data constitutes of thousands

of features. Verily selection process diminishes the number of features by removing the irrelevant andnoisy factors and thus makes the complete investigations more feasible, methodical and canonical[11]. The pivotal disbenefit of FS is that it is not ideal for homogeneous data. FS when applied to homogeneous datasets resulted in variability in stability[3].So proposed estimations are Q-statistic[5] and Booster with a classifier respectively, which consolidates the stability of the features. Proposed system not only provides the high forecast model but also stability is achieved. The complications with the existing system and dominance of the proposed systems are discussed in this paper.

Existing system:

- One regularly utilized approach[18] is to first discretize the consistent an outstanding in the preprocessing step and utilize shared data (MI)[9] to choose significant highlights.
- This is on account of finding important highlights in view of the discretized MI[9] are moderately straightforward.
- When finding the correct[11] suitable features especially from the unlimited records.
- These records are with high consistency

through utilizing the persistent data is an impressive procedure[20].

- Several examines in view of resampling[15] strategy have been done to produce distinctive informational indexes.
- For arrangement issue and a portion of the investigations use resampling on the element space.
- The motivations behind every one of these investigations are on the forecast precision of grouping without thought on the solidness of the chosen highlighted subset.

Drawbacks of Existing system:

- Majority of the effective FS algorithm[1] in multi-dimensional issues have used forward choice technique yet not considered in reverse end strategy since it is illogical to execute in reverse end process with gigantic number of highlights.
- A genuine inborn issue with forward determination is, nonetheless, a flip in the choice of the underlying component may prompt a totally extraordinary element subset and thus the security of the chosen include set will be low in spite of the fact that the choice may yield high precision[9].
- Devising a productive strategy to acquire a more steady component subset with high precision is a testing territory of research.

Proposed system:

- This paper suggests a Q-statistic[4] to assess the execution of a FS with at least one classifier.[13][14].
- It is a mixed calculative measure[5] of the forecast precision of the classifier and the dependability at that

specific point.

- Proposes performance booster on the choice within subset from a given FS algorithm.
- The fundamental thought of boosting an application is to acquire a few informational collections from unique informational index by resampling[7] on test space.
- At that point FS algorithm is connected to each of these resampled informational indexes[7][12] to acquire diverse element subsets.
- The combination of subsets will be the component subset got by the Booster of FS algorithm.

Advantages of Proposed system:

- Empirical thinks about demonstrate that the Booster of a calculation helps the estimation of Q-statistic[9] as well as the expectation exactness of the classifier connected.
- Particularly, the execution of mRMR-Booster[19] was appeared to be remarkable both in the changes of forecast precision and Q-statistic[4].

1. FEATURE SELECTION:

Feature Selection[1] is an algorithm which takes the dataset as input, and performs its operations on it. The properties in the database is called as features and the algorithm selects the features for the further proceedings like redundancy check etc., is called as feature selection. Without feature selection[1] there is no work done on the dataset. When the patient tries to enter the redundant data, then the feature are checked with the already existing features and fulfills the request. If the features are matched then it will say that it is redundant data otherwise the application will enter the patient details into the database. There are 6500 datasets included in the project. [14] 17

Features includes provider id, Hospital name, Address, city, state, measure starting date, ending date, measure name, phone number, Compared national, Denominator, Score, Lower estimate, Higher estimate, and Measure id. The aim of project is to find the death rate of patients in the respective hospitals.

3. Methodology :

In methodology the workflow of the project going to be discussed. Here, the description of the following steps are:[1][11][14][15].

- Firstly, starting the process,
- Loading the 6500 datasets.
- In the 3rd step if any duplication of data found then it is removed.
- Feature Selection has two lay-offs mainly **Forward Selection** and **Backward Elimination**.
- Forward Selection adds on the data where as it results in dimensionality [5] problem. On the other side removing of features is such a problematic task and not possible with Backward Elimination.
- Then strong redundancy[22] check is done .In this step it de duplicates the data completely.
- Data gets classified and finally evaluated feature selection is obtained.
- It indeed results in accuracy.

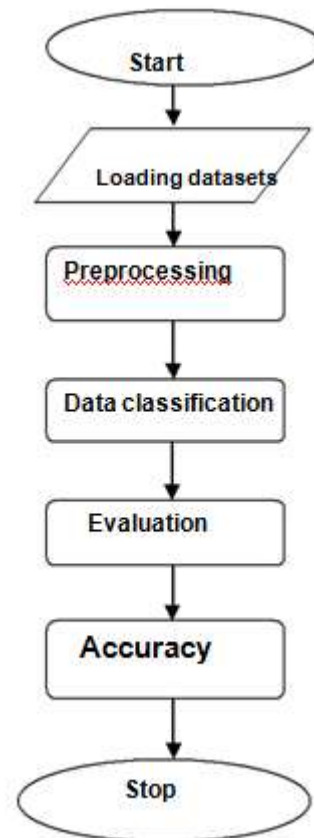


Fig 3.1 Workflow of the process

4. Implementation:

In this paper Booster algorithm used for successful execution of the project. And based on the Booster algorithm the testing and associated results are carried out.

Modules:

- Dataset Collection
- Feature Selection
- Removing Irrelevant Features
- Booster accuracy.

Modules Description:

-Dataset Collection:

To gather as well as recover information about exercises, results, setting and different variables. And the information is stored in the database.

-Feature Selection:

This[1] is a required combination measure of the forecast exactness of the classifier and the dependability of the chose queried data. At that point the paper proposes Booster on the determination of highlight subset from a given FS calculation. FS in high dimensional information

needs preprocessing procedure to choose just significant highlights or to sift through superfluous highlights.

-Removing Irrelevant Features:

The irrelevant features[11] are removed during the preprocessing step. The irrelevant features[7][11] in this project are entry of multiple records.

-Booster accuracy:

The Booster of a FS calculation that lifts the estimation of the Q-statistic of the calculation connected. Exact examinations in light of manufactured information Empirical investigations [19] demonstrate that the Booster of a calculation supports the estimation of Q-measurement as well as the forecast exactness of the classifier connected. The assessment of the relative execution for the effectiveness of s-Booster over the first FS calculations in view of the forecast exactness and Q-statistic.two Boosters, FAST-Booster, FCBF-Booster and mRMR-Booster. mRMR-Booster[9] enhances exactness extensively: general normal precision. One fascinating point to note here is that mRMR-Booster is more effective in boosting the exactness .The FAST-Booster likewise enhances precision, yet not as high as mRMR.

ALGORITHM:

Booster Algorithm: Booster b

Input: FS algorithm + Data Set + total number of partitions.

Output: Feature subset

selected is V^* 1. Split D into partitions

2. $V^*=0$

3. for $i=1$ to b do

4. $D_{-i} = D - D_i$ # remove D_i from

5. $V_{-i} <- s(D_{-i})$ # obtain V_{-i} by applying s on D_{-i}

6. $V^*=V^* \cup V_i$

7. end for

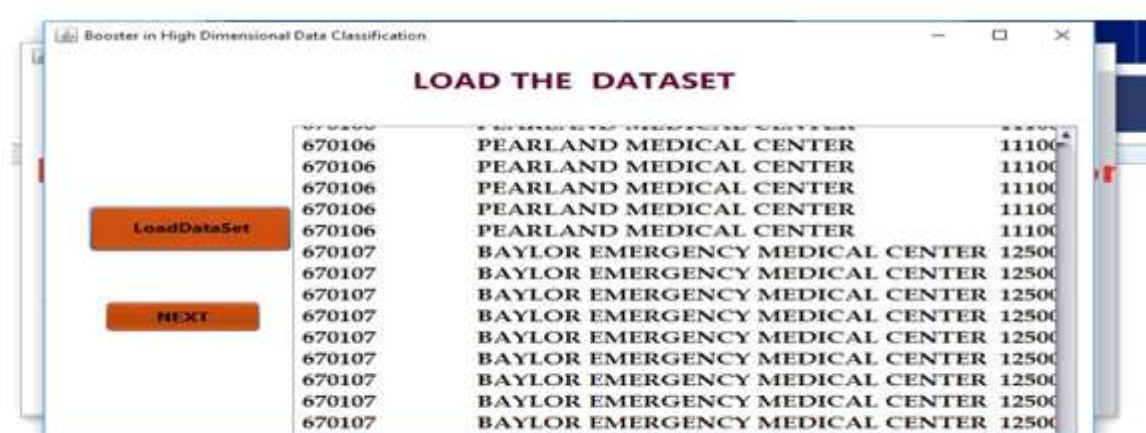
8. return V^*

The workflow of the algorithm starts as,

- The whole data is divided into partitions.
- If any duplication occurs then eliminates.
- Then the strong redundancy check is carried out.
- If any inconsistencies then removed in 3rd step and Process ends.

5. Table and results :

In this session, according to the project there are about 6500 datasets and 17 features. Out of these only 3 features and 10 datasets are illustrated in the paper.[17]



Dataset ID	Hospital Name	Value
670106	PEARLAND MEDICAL CENTER	11100
670106	PEARLAND MEDICAL CENTER	11100
670106	PEARLAND MEDICAL CENTER	11100
670106	PEARLAND MEDICAL CENTER	11100
670106	PEARLAND MEDICAL CENTER	11100
670106	PEARLAND MEDICAL CENTER	11100
670107	BAYLOR EMERGENCY MEDICAL CENTER	12500
670107	BAYLOR EMERGENCY MEDICAL CENTER	12500
670107	BAYLOR EMERGENCY MEDICAL CENTER	12500
670107	BAYLOR EMERGENCY MEDICAL CENTER	12500
670107	BAYLOR EMERGENCY MEDICAL CENTER	12500
670107	BAYLOR EMERGENCY MEDICAL CENTER	12500
670107	BAYLOR EMERGENCY MEDICAL CENTER	12500
670107	BAYLOR EMERGENCY MEDICAL CENTER	12500

Fig 5.1 Loading of datasets.

- In the existing system, there will be no strong redundancy check.[21]

- After the split of datasets new records are added.
- In the proposed system, Q-statistic is proposed to lift up the execution of the project with classifiers.
- It forecasts the stability and eliminates any variations in features.
- Then proposed Booster a technique, to re-test the sample space.[5]
- Booster results in a very strict , erect strong redundancy check.[22]
- Three algorithms FCBF,FAST, mRmR are being used. Out of the three best proven algorithm is mRmR.
- These algorithms are worked implicitly.



Fig 5.2 Feature selection with classification.

The data is classified according to the algorithm and it is shown to the user in a convenient way.

- Feature Selection[1][16] aims in minimizes redundancy and maximizes relevant target.
- FS follows one disadvantageous step in checking: **Data classification without redundancy.**
- Finding redundant data is important. As in Feature selection[1] algorithm classification is done without a proper redundancy check in database which in turn results in space complexity[17] in memory. The last step is the evaluation step .The Booster algorithm[15] and Q-statistic here is applied on the datasets. And removes the redundant data to the maximum extent.



Fig 5.3 Evaluation process

The three algorithms FCBF ,mRmR, FAST[12] works internally .The main aim of the project is to estimate the highest and lowest death rates in

hospitals. The below tables are results in comparison and accuracy then to previous hospitals.

Provider_id	Hospital_name	Compared_national	Score
10001	Memorial hospital	Acute Myocardial Infarction	1657.5
10011	St.Joseph's hospital	Heart failure	1292.90
11020	Good Samaritan	Pneumonia (PN) 30-Day Mortality Rate	1212.80
11035	St.Francis	Death rate for chronic obstructive pulmonary disease	100.09
12045	Memorial Medical	Death rate for CABG	866.20
12050	Mercy medical center	Lung disease	811.0
13123	Doctor's Hospital	Rate of unplanned readmission for CABG	686.099
15231	Cottage hospital	Acute Myocardial Infarction	99.5
16789	Delaware valley	Death rate for chronic obstructive pulmonary disease	110.7
23045	Mercy hospital	Infectious disease.	801.5

Table 5.1.1 Highest death survey.

Provider_id	Hospital_name	Compared_national	Score
1435	West Covina Medical Center	Acute Myocardial Infarction	4.8
1537	Turnining point	Heart failure	12.0
1629	Hoag Orthopedic Institute	Pneumonia (PN) 30-Day Mortality Rate	14.1
1700	Pagosa Springs Medical center	Death rate for chronic obstructive pulmonary disease	14.0000
12046	Fairbanks	Death rate for CABG	14.699
13050	Kaiser Foundation Hospital	Lung disease	13.3
14123	Treasure Valley Hospital	Rate of unplanned readmission for CABG	2.8
15231	Pawnee Valley community	Acute Myocardial Infarction	1.5
16789	Columbia Medical center	Death rate for chronic obstructive pulmonary disease	1.7
23099	Claiborne County Hospital	Infectious disease.	1.0

Table 5.1.2 Lowest death survey

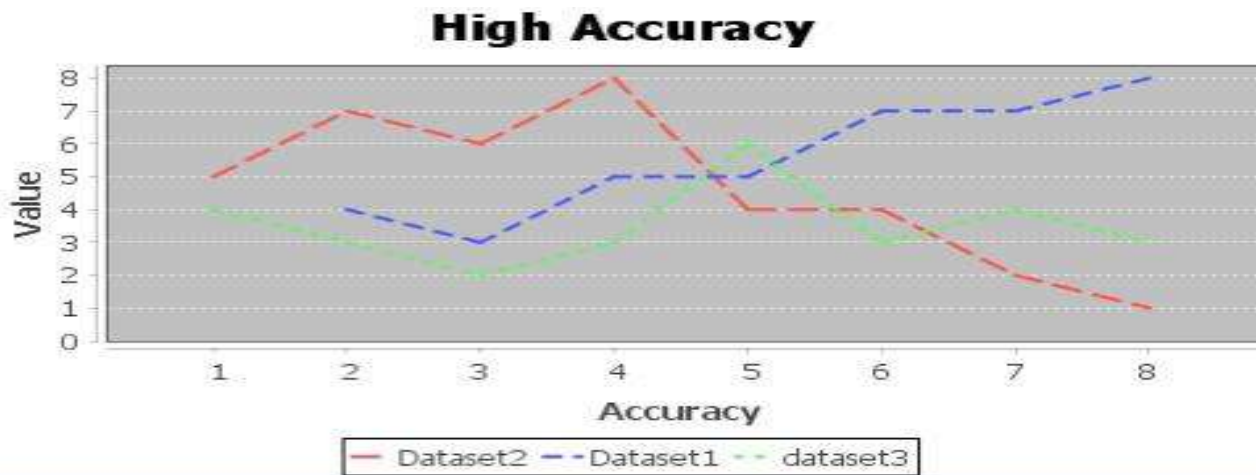


Fig 5.1.3 Accuracy graph.

This graph gives the clear information about the result. On the x-axis takes the values and on the Y-axis the performance is taken. This is final outcome of the project, which explains about the accuracy of different inputs.

6. CONCLUSION:

The proposed a measure Q-statistic[2] assesses the execution of a FS calculation. Q-statistic accounts both for the solidness of chose include subset and the forecast exactness[16]. The paper proposed Booster to support the execution of a current FS calculation. Experimentation have demonstrated successfully and recommends Booster as sit enhances the forecast exactness and the Q-statistic[5] of the three understood Fs particularly, Booster was appeared to exceptional both in the upgrades of expectation exactness and Q-statistic. It was watched that if a FS algorithm[1] is proficient however couldn't acquire superior in the exactness or the Q-statistic for some particular information, Booster of the FS calculation will support the execution. In any case, if a FS algorithm itself is not productive, Booster will most likely be unable to acquire superior. The execution of Booster relies upon the execution of the FS calculation connected.

7. REFERENCES

- [1] K.M .Ting, J.R. Wells,S.C Tan, S.W.Teng, and G.I Webb,"Feature –subspace aggregating: Ensembles for stable and unstable learners,"*Mach.Learn.*,vol.82, no.3,pp.375-397,2011.
- [2] D. Aha and D. Kibler, "Instance-based learning algorithms," *Mach. Learn.*, vol. 6, no. 1, pp. 37–66, 1991.
- [3] S. Alelyan, "On feature selection stability: A data perspective," PhD dissertation, Arizona State Univ., Tempe, AZ, USA, 2013.
- [4] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. M. Izidore, S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. H. Jr, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.
- [5] L.I.Kuncheva, "A stability index for feature selection," in *Proc Artif. Intell. Appl.*, pp. 421–427, 2007.
- [6] F. Alonso-Atienza, J. L. Rojo-Alvarez, A. Rosado-Muñoz, J. J. Vinagre, A. Garcia-Alberola, and G. Camps-Valls, "Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection," *Expert Syst. Appl.*, vol. 39, no. 2, pp. 1956–1967, 2012.
- [7] P. J. Bickel and E. Levina, "Some theory for Fisher's linear discriminant function, naive Bayes, and some alternatives when there are many more variables than observations," *Bernoulli*, vol. 10, no. 6, pp. 989–1010, 2004.
- [8] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, "Kernel density estimation via diffusion,"



- The
Ann. Statist., vol. 38, no. 5, pp. 2916–2957, 2010.
- [9] G. Brown, A. Pocock, M. J. Zhao, and M. Lujan, “Conditional likelihood maximization: A unifying framework for information theoretic feature selection,” *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 27–66, 2012.
- [10] C. Kamath, *Scientific data mining: a practical perspective*, Siam, 2009.
- [11] G.H.John,R.Kohavi,andK.Pfleger, “Irrelevant features and the subset selection problem”,in *Proc.11th Int.Conf.Mach.Learn.*,vol.94, pp.121-129, 1994.
- [12] C. Corinna and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
- [13] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Series in Telecommunications and Signal Processing)*, 2nd ed. Hoboken, NJ, USA: Wiley, 2002.
- [14] D. Dembele, “A flexible microarray data simulataion model,” *Microarrays*, vol. 2, no. 2, pp. 115–130, 2013.
- [15] D. Derroncourt, B. Hanczar, and J. D. Zucker, “Analysis of feature selection stability on high dimension and small sample data,” *Comput. Statist. Data Anal.*, vol. 71, pp. 681–693, 2014.
- [16] J. Fan and Y. Fan, “High dimensional classification using features annealed independence rules,” *Ann. Statist.*, vol. 36, no. 6, pp. 2605–2637, 2008.
- [17] J. Fan, P. Hall, and Q. Yao, “To how many simultaneous hypothesis tests can normal, Student’s *t* or bootstrap calibration be applied?,” *J. Am. Statist. Assoc.*, vol. 102, no. 480, pp. 1282–1288, 2007.
- [18] U. M. Fayyad and K. B. Irani, “Multi-interval discretization of continuous-valued attributes for classification learning,” *Artif. Intell.*, vol. 13, no. 2, pp. 1022–1027, 1993.
- [19] A. J. Ferreira and M. A. T. Figueiredo, “Efficient feature selection filters for high dimensional data,” *Pattern Recog. Lett.*, vol. 33, no. 13, pp. 1794–1804, 2012.
- [20] B. Franay, G. Doquire, and M. Verleysen, “Theoretical and empirical study on the potential inadequacy of mutual information for feature selection in classification,” *Neurocomputing*, vol. 112, pp. 64–78, 2013.
- [21] H.Peng,F.Long,C.Ding, “Feature selection based on mutual information criteria of max dependency, max relevance, andminredundancy ”*IEEETrans.Pattern.Anal.Mach Intell.*, vol.27, no.8, pp. 1226– 1238, Aug. 2005.
- [22] L.Yu and H.Liu, “Efficient feature selection via analysis of relevance and redundancy,”*The J.Mach.Learn.Res.*,vol.5,no.2,pp.1205-1224,2004.