ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

A New Evaluation Measure Q-Statistic for Stability of the Selected Feature Subset and Prediction Accuracy

V. Shamsheer[1] , A.Chandra Sekhar[2], A. L. Sreenivasulu[3]

M.Tech Student, Dept. of CSE, GATES Engineering College, Affiliated to JNTUA , Andhra Pradesh, India[1]

Assistant Professor, Dept. of CSE, GATES Engineering College, Affiliated to JNTUA, Andhra Pradesh, India[2]

Associate Professor & HOD, Dept. of CSE, GATES Engineering College, Affiliated to JNTUA, Andhra Pradesh, India[3]

ABSTRACT: In present, duplicate detection methods need to process ever larger datasets in ever shorter time, maintaining the quality of a dataset becomes increasingly difficult. Feature selection plays a significant role in improving the performance of the machine learning algorithms in terms of reducing the time to build the learning model and increasing the accuracy in the learning process. Therefore, the researchers pay more attention on the feature selection to enhance the performance of the machine learning algorithms. Identifying the suitable feature selection method is very essential for a given machine learning task with high-dimensional data. Hence, it is required to conduct the study on the various feature selection methods for the research community especially dedicated to develop the suitable feature selection method for enhancing the performance of the machine learning tasks on high-dimensional data. In order to fulfill this objective, this paper devotes the complete literature review on the various feature selection methods for high-dimensional data..

1. INTRODUCTION

Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Data mining derives

its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides. Although data mining is still in its infancy, companies in a wide range of industries - including retail, finance, healthcare, manufacturing transportation, and aerospace - are already using data mining tools and techniques to take advantage of historical data. By using pattern recognition technologies and statistical and mathematical techniques to sift through warehoused information, data mining helps analysts recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed. For businesses, data mining is used to discover
patterns and relationships in the data in order to help make better business decisions. Data mining can help spot sales trends, develop smarter marketing campaigns, and accurately predict customer loyalty. Data mining technology can generate new business opportunities by:

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

1.1.1Automated Prediction Of Trends And Behaviors:

Data mining automates the process of finding predictive information in a large database. Questions that traditionally required extensive hands-on analysis can now be directly answered from the data. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

1.1.2Automated discovery of previously unknown patterns:

Data mining tools sweep through databases and identify previously hidden patterns. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors. While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

II. EXISTING SYSTEM

- One often used approach is to first discretised the continuous features in the preprocessing step and use mutual information (MI) to select relevant features. This is because finding relevant features based on the discretised MI is relatively simple while finding relevant features directly from a huge number of the features with continuous values using the definition of relevancy is quite a formidable task.
- Several studies based on resampling technique have been done to generate different data sets for classification problem and some of the studies utilize resampling on the feature space.
- The purposes of all these studies are on the prediction accuracy of classification without consideration on the stability of the selected feature subset.

DISADVANTAGES OF EXISTING SYSTEM:

- Most of the successful FS algorithms in high dimensional problems have utilized forward selection method but not considered backward elimination method since it is impractical to implement backward elimination process with huge number of features.
- A serious intrinsic problem with forward selection is, however, a flip in the decision of the initial feature may lead to a completely different feature subset and hence the stability of the selected feature set will be very low although the selection may yield very high accuracy.
- Devising an efficient method to obtain a more stable feature subset with high accuracy is a challenging area of research.

III. PROPOSED SYSTEM

- This paper proposes Q-statistic to evaluate the performance of an FS algorithm with a classifier. This is a hybrid measure of the prediction accuracy of the classifier and the stability of the selected features. Then the paper proposes Booster on the selection of feature subset from a given FS algorithm.
- The basic idea of Booster is to obtain several data sets from original data set by resampling on sample space. Then FS algorithm is applied to each of these resampled data sets to obtain different feature subsets. The union of these selected subsets will be the feature subset obtained by the Booster of FS algorithm.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

ADVANTAGES OF PROPOSED SYSTEM:

- Empirical studies show that the Booster of an algorithm boosts not only the value of Q-statistic but also the prediction accuracy of the classifier applied.
- We have noted that the classification methods applied to Booster do not have much impact on prediction accuracy and Q-statistic. Especially, the performance of mRMR-Booster was shown to be outstanding both in the improvements of prediction accuracy and Q-statistic.

IV. MODULE DESCRIPTION

A. Dataset Collection

To collect and/or retrieve information concerning activities, results, context and alternative factors. it's vital to contemplate the sort of data it need to assemble from your participants and therefore the ways in which you may analyze that information. the information set corresponds to the contents of one info table, or one applied math information matrix, wherever each column of the table represents a specific variable. when aggregation the information to store the info**.**

B. Preprocessing Method

Data Preprocessing or information improvement, information is cleaned through processes like filling in missing values, smoothing the wheezy information, or resolution the inconsistencies within the information. And additionally accustomed removing the unwanted information. ordinarily used as a preliminary data processing follow, information preprocessing transforms the information into a format which will be a lot of simply and effectively processed for the aim of the user.

C. Data Separation

After finishing the preprocessing, the information separation to be performed. The block algorithms assign every record to a set cluster of comparable records (the blocks) and so compare all pairs of records inside these groups. every block inside the block comparison matrix represents the comparisons of all records in one block with all records in another block, the equal block, all blocks have constant size.

D. Duplicate Detection

The duplicate detection rules set by the administrator, the system alerts the user concerning potential duplicates once the user tries to form new records or update existing records. to keep up information quality, you'll be able to schedule a replica detection job to examine for duplicates for all records that match a particular criteria. you'll be able to clean the information by deleting, deactivating, or merging the duplicates rumored by a replica detection

E. Quality Measures

The quality of those systems is, hence, measured employing a cost-benefit calculation. particularly for ancient duplicate detection processes, it's troublesome to satisfy a budget limitation, as a result of their runtime is difficult to predict. By delivering as several duplicates as potential in an exceedingly given quantity of your time, progressive processes optimize the cost-benefit quantitative relation. In producing, a live of excellence or a state of being free from defects, deficiencies and vital variations. it's caused by strict and consistent commitment to sure standards that bring home the bacon uniformity of a product so as to satisfy specific client or user needs.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

V. CONCLUSION

This paper proposed a measure Q-statistic that evaluates the performance of an FS algorithm. Q-statistic accounts both for the stability of selected feature subset and the prediction accuracy. Also we have noted that the classification methods applied to Booster do not have much impact on prediction accuracy and Q-statistic. Especially, the performance of m RMR-Booster was shown to be outstanding both in the improvements of prediction accuracy and Qstatistic. It was observed that if an FS algorithm is efficient but could not obtain high performance in the accuracy or the Q-statistic for some specific data, Booster of the FS algorithm will boost the performance. However, if an FS algorithm itself is not efficient, Booster may not be able to obtain high performance. The performance of Booster depends on the performance of the FS algorithm applied. If Booster does not provide high performance, it implies two possibilities: the data set is intrinsically difficult to predict or the FS algorithm applied is not efficient with the specific data set. Hence, Booster can also be used as a criterion to evaluate the performance of an FS algorithm or to evaluate the difficulty of a data set for classification.

REFERENCES

[1] S. E. Whang, D. Marmaros, and H. Garcia-Molina, **"**Pay-as-you-go entity resolution,**"** IEEE Trans. Knowl. Data Eng., vol. 25, no. 5,pp. 1111–1124, May 2012.
[2]. A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, **"**Duplicate record detection: A survey,**"** IEEE Trans. Knowl. Data Eng., vol. 19,no. 1, pp. 1–16, Jan. 2007.
[3]. F. Naumann and M. Herschel, An Introduction to Duplicate Detection.San Rafael, CA, USA: Morgan & Claypool, 2010.
[4]. H. B. Newcombe and J. M. Kennedy, **"**Record linkage: Making maximum use of the discriminating power of identifying information,**"** Commun. ACM, vol. 5, no. 11, pp. 563–566, 1962.
[5]. M. A. Hernandez and S. J. Stolfo, **"**Real-world data is dirty: Data cleansing and the merge/purge problem,**"** Data Mining Knowl. Discovery, vol. 2, no. 1, pp. 9–37, 1998.
[6]. X. Dong, A. Halevy, and J. Madhavan, **"**Reference reconciliation in complex information spaces,**"** in Proc. Int. Conf. Manage. Data,2005, pp. 85–96.
[7]O. Hassanzadeh, F. Chiang, H. C. Lee, and R. J. Miller,**"**Framework for evaluating clustering algorithms in duplicate detection,**"** Proc. Very Large Databases Endowment, vol. 2, pp. 1282–1293, 2009.
[8]O. Hassanzadeh and R. J. Miller, **"**Creating probabilistic databases from duplicated data,**"** VLDB J., vol. 18, no. 5, pp. 1141–1166, 2009.
[9]U. Draisbach, F. Naumann, S. Szott, and O. Wonneberg,**"**Adaptive windows for duplicate detection,**"** in Proc. IEEE 28th Int. Conf. Data Eng., 2012, pp. 1073–1083.
[10] S. Yan, D. Lee, M.-Y. Kan, and L. C. Giles, **"**Adaptive sorted neighborhood methods for efficient record linkage,**"** in Proc. 7th ACM/IEEE Joint Int. Conf. Digit. Libraries, 2007, pp. 185–194.
[11] J. Madhavan, S. R. Jeffery, S. Cohen, X. Dong, D. Ko, C. Yu, and A. Halevy, **"**Web-scale data integration: You can only afford to pay as you go,**"** in Proc. Conf. Innovative Data Syst.Res., 2007.
[12] S. R. Jeffery, M. J. Franklin, and A. Y. Halevy, **"**Pay-as-you-go user feedback for dataspace systems,**"** in Proc. Int. Conf. Manage. Data,2008, pp. 847–860.