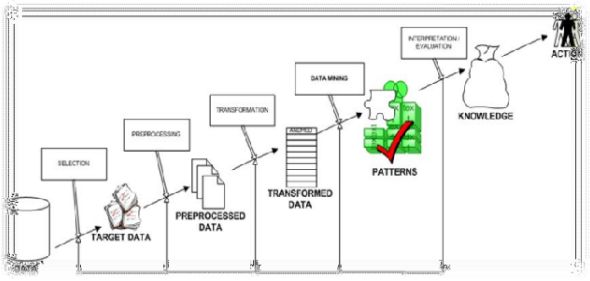
BOOSTER IN HIGH DIMENSIONAL DATA CLASSIFICATION

[1]Dr C.Suresh Kumar, [2]Dr R.Anand, [3]Dr Nedunchezhian, [4]Dr Prabakaran
1,2,3,4 Professor, Department of Computer Science and Engineering,
Malla Reddy College of Engineering, Hyderabad

Abstract—Classification problems in high dimensional data with a small number of observations are becoming more common especially in microarray data. During the last two decades, lots of efficient classification models and feature selection (FS) algorithms have been proposed for higher prediction accuracies. However, the result of an FS algorithm based on the prediction accuracy will be unstable over the variations in the training set, especially in high dimensional data. This paper proposes a new evaluation measure Q-statistic that in corporate the stability of the selected feature subset in addition to the prediction accuracy. Then, we propose the Booster of an FS algorithm that boosts the value of the Q- statistic of the algorithm applied. Empirical studies based on synthetic data and 14 micro array datasets show that Booster boosts not only the value of the Q-statistic but also the prediction accuracy of the algorithm applied unless the dataset is intrinsically difficult to predict with the given algorithm.

INTRODUCTION



Structure of Data Mining

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases..

Data mining consists of five major elements:

1) Extract, transform, and load transaction data on to the data ware house system.

2) Store and manage the data in a multidimensional database system.

3) Provide data access to business analysts and information technology professionals.

4) Analyze the data by application software.

5) Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

• Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

• Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

• Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) data set to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi- way splits. CART typically requires less data

ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME-4, ISSUE-11, 2017
DOI: 10.21276/ijcesr.2017.4.11.45

preparation than CHAID.

• Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the

classes of the k record(s) most similar to it in a historical dataset (where k=1). Sometimes called the k-nearest neighbor technique.

• Rule induction: The extraction of useful if-then rules from data based on statistical significance.

• Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

LITERATURE SURVEY

Biomarker discovery is an important topic in biomedical applications of computational biology, including applications such as gene and SNP selection from high- dimensional data. However, robustness of biomarkers is an important issue, as it may greatly influence subsequent biological validations. First contribution is a general framework for the analysis of the robustness of a biomarker selection algorithm. Secondly, they conducted a large-scale analysis of the recently introduced concept of ensemble feature selection, where multiple feature selections are combined in order to increase the robustness of the final set of selected features. We focus on selection methods that are embedded in the estimation of support vector machines (SVMs). SVMs are powerful classification models that have shown state-of-the-art performance on several diagnosis and prognosis tasks on biological data. Their feature selection extensions also offered good results for gene selection tasks. they show that the robustness of SVMs for biomarker discovery can be substantially increased by using ensemble feature selection techniques, while at the same time improving upon classification performances. The proposed methodology is evaluated on four microarray datasets showing increases of up to almost 30% in robustness of the selected biomarkers, along with an improvement of ~15% in classification performance. The stability improvement with ensemble methods is particularly noticeable for small signature sizes (a few tens of genes), which is most relevant for the design of a diagnosis or prognosis model from a gene signature[1].

Storing and using specific instances improves the performance of several supervised learning algorithms. These include algorithms that learn decision trees, classification rules, and distributed networks. However, no investigation has analyzed algorithms that use only specific instances to solve incremental learning tasks. In this paper, we describe a framework and methodology, called instance- based learning, that generates classification predictions using only specific instances. Instance-based learning algorithms do not maintain a set of abstractions derived from specific instances. This approach extends the nearest neighbor algorithm, which has large storage requirements. We describe how storage requirements can be significantly reduced with, at most, minor sacrifices in learning rate and classification accuracy. While the storage-reducing algorithm performs well on several real-world databases, its performance degrades rapidly with the level of attribute noise in training instances. Therefore, we extended it with a significance test to distinguish noisy instances. This extended algorithm's performance degrades gracefully with increasing noise levels and compares favorably with a noise tolerant decision tree algorithm[2].

Diffuse large B-cell lymphoma (DLBCL), the most common subtype of non-Hodgkin's lymphoma, is clinically heterogeneous: 40% of patients respond well to current therapy and have prolonged survival, whereas the remainder succumb to the disease. [3] proposed that this variability in natural history reflects unrecognized molecular heterogeneity in the tumours. Using DNA microarrays, They have conducted a systematic characterization of gene expression in B-cell malignancies and show that there is diversity in gene expression among the tumours of DLBCL patients, apparently reflecting the variation in tumour proliferation rate, host response and differentiation state of the tumour. They identified two molecularly distinct forms of DLBCL which had gene expression patterns indicative of different stages of B-cell differentiation. One type expressed genes characteristic of germinal centre B cells ('germinal centre B-like DLBCL'); the second type expressed genes normally induced during in vitro activation of peripheral blood B cells ('activated B-like DLBCL').

ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME-4, ISSUE-11, 2017
DOI: 10.21276/ijcesr.2017.4.11.45

Patients with germinal centre B-like DLBCL had a significantly better overall survival than those with activated B-like DLBCL. The molecular classification of tumours on the basis of gene expression can thus identify previously undetected and clinically significant subtypes of cancer.

Oligo nucleotide arrays can provide a broad picture of the state of the cell, by monitoring the expression level of thousands of genes at the same time. It is of interest to develop techniques for extracting useful information from the resulting data sets. [4] report the application of a two- way clustering method for analyzing a data set consisting of the expression patterns of different cell types. Gene expression in 40 tumor and 22 normal colon tissue samples was analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes. An efficient two-way clustering algorithm was applied to both the genes and the tissues, revealing broad coherent patterns that suggest a high degree of organization underlying gene expression in these tissues. Coregulated families of genes clustered together, as demonstrated for the ribosomal proteins. Clustering also separated cancerous from noncancerous tissue and cell lines from in vivo tissues on the basis of subtle distributed patterns of genes even when expression of individual genes varied only slightly between the tissues. Two-way clustering thus may be of use both in classifying genes into functional groups and in classifying tissues based on geneexpression.

Early detection of ventricular fibrillation (VF) is crucial for the success of the defibrillation therapy in automatic devices. A high number of detectors have been proposed in [5] based on temporal, spectral, and time– frequency parameters extracted from the surface electrocardiogram (ECG), showing always a limited performance. The combination ECG parameters on different domain (time, frequency, and time–frequency) using machine learning algorithms has been used to improve detection efficiency. However, the potential utilization of a wide number of parameters benefiting machine learning schemes has raised the need of efficient feature selection (FS) procedures. In this study, we propose a novel FS algorithm based on support vector machines (SVM) classifiers and bootstrap re sampling (BR) techniques. We define a backward FS procedure that relies on evaluating changes in SVM performance when removing features from the input space. This evaluation is achieved according to a nonparametric statistic based on BR. After simulation studies, we benchmark the performance of our FS algorithm in AHA and MIT-BIH ECG databases. Our results show that the proposed FS algorithm outperforms the recursive feature elimination method in synthetic examples, and that the VF detector performance improves with the reduced feature set.

IMPLEMENTATION

MODULES:

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

MODULES DESCSRIPTION:

Dataset Collection:

To collect and/or retrieve data about activities, results, context and other factors. It is important to consider the type of information it want to gather from your participants and the ways you will analyze that information. The data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable. after collecting the data to store the Database.

Feature Selection:

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 data sets from original data set by re sampling on sample space. Then FS algorithm is applied to each of these re sampled 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. the Booster of an algorithm boosts not only the value of Q-statistic but also the prediction accuracy of the classifier applied. Several studies based on re sampling technique have been done to generate different data sets for classification problem , and some of the studies utilize re sampling on the feature space . The purposes of all these studies are on the prediction accuracy of classification without

ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME-4, ISSUE-11, 2017
DOI: 10.21276/ijcesr.2017.4.11.45

consideration on the stability of the selected feature subset. FS algorithms— FAST, FCBF, and mRMR—and their corresponding Boosters, we apply k-fold cross validation. For this, k training sets and their corresponding k test sets are generated. For each training set, Booster is applied to obtain V . Classification is performed based on the training set with the selection V , and the test set is used for prediction accuracy

Removing Irrelevant Features:

The features of high dimensional microarray data are irrelevant to the target feature and the proportion of relevant features or the percentage of up-regulated Finding relevant features simplifies learning process and increases prediction accuracy. The finding, however, should be relatively robust to the variations in training data, especially in biomedical study, since domain experts will invest considerable time and efforts on this small set of selected features. The pre- processing steps to find weakly relevant features based on t- test and to remove irrelevant features based on MI.FS in high dimensional data needs preprocessing process to select only relevant features or to filter out irrelevant features. the selected subsets V1; . . . ; Vb obtained by s consist only of the relevant features where redundancies are removed, V will include more relevant features where redundancies are removed. Hence, V will induce smaller error of selecting irrelevant features. However, if s does not completely remove redundancies, V may result in the accumulation of larger size of redundant features. find more relevant features but may include more irrelevant features, and also may induce more redundant features. This is because no FS algorithm can select all relevant features while removing all irrelevant features and redundant features.

Booster accuracy:

The Booster of an FS algorithm that boosts the value of the Q-statistic of the algorithm applied. Empirical studies based on synthetic data 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. Booster is simply a union of feature subsets obtained by a resembling technique. The resembling is done on the sample space. Booster needs an FS algorithm s and the number of partitions b. When s and b are needed to be specified, we will use notation s-Booster. Hence, s-Booster1 is equal to s since no partitioning is done in this case and the whole data is used. When s selects relevant features while removing redundancies, s-Booster will also select relevant features while removing redundancies. the notation FAST-Booster, FCBF-Booster, and mRMR-Booster for the Booster of the corresponding FS algorithm. we will evaluate the relative performance efficiency of s-Booster over the original FS algorithm s based on the prediction accuracy and Q- statistic. two Boosters, FAST-Booster , FCBF-Booster and mRMR-Booster. mRMR- Booster improves accuracy considerably: overall average accuracy. One interesting point to note here is that mRMR-Booster is more efficient in boosting the accuracy .we can observe that FAST-Booster also improves accuracy, but not as high asmRMR

SYSTEM ANALYSIS

EXISTING SYSTEM:

- One often used approach is to first discretize 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 discretized 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 re sampling technique have been done to generate different data sets for classification problem and some of the studies utilize re sampling 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 features subset.

DISADVANTAGES OF EXISTING SYSTEM:

- Most of the successful FS algorithms in high dimensional problems have utilized forward selection method but not considered backward limination 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

ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME-4, ISSUE-11, 2017
DOI: 10.21276/ijcesr.2017.4.11.45

yield very high accuracy.

☐ Devising an efficient method to obtain a more stable feature subset with high accuracy is a challenging area of research.

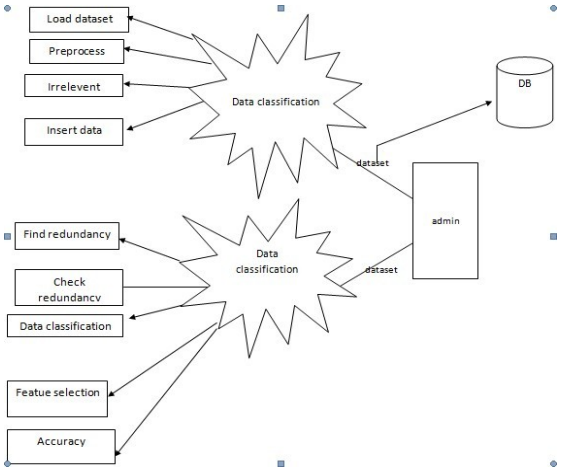
PROPOSED SYSTEM:

☐ 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 re sampling on sample space. Then FS algorithm is applied to each of these re sampled 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.

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

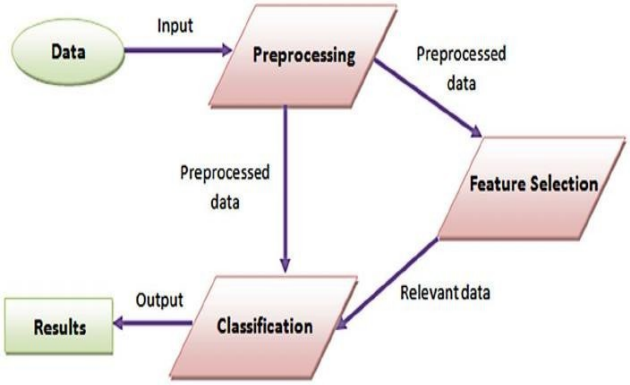
Information flow and the transformations that are applied as data moves from input to out put.

4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

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.

SYSTEM DESIGN

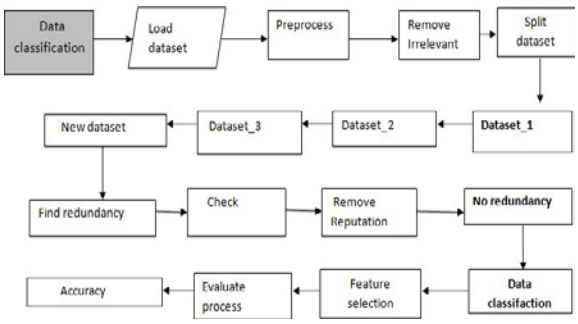
SYSTEM ARCHITECTURE:



BLOCK DIAGRAM:

DATA FLOW DIAGRAM:

1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
2. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts



ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME-4, ISSUE-11, 2017
DOI: 10.21276/ijcesr.2017.4.11.45

SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are technologies used are freely available. Only the customized products had to be purchased.

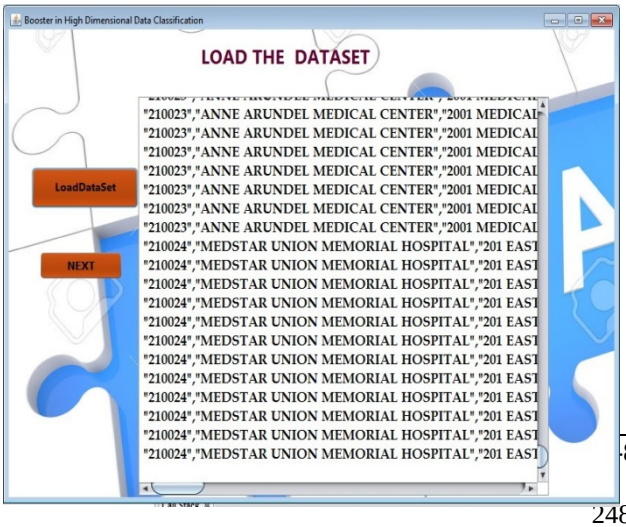
TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

SCREEN SHOTS



OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.
2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.
3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.
2. Select methods for presenting information.
3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following

objectives.

☐ Convey information about past activities, current status or projections of the

☐ Future.

☐ Signal important events, opportunities, problems, or warnings.

☐ Trigger anaction.

☐ Confirm anaction.

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. The paper proposed Booster to boost the performance of an existing FS algorithm. Experimentation with synthetic data and 14 microarray data sets has shown that the suggested Booster improves the prediction accuracy and the Q-statistic of the three well-known FS algorithms: FAST, FCBF, and mRMR. 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 mRMR-Booster was shown to be outstanding both in the improvements of prediction accuracy andQ-statistic.

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: Dataset is intrinsically difficult to predictor 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] T. Abeel, T. Helleputte, Y. V. de Peer, P. Dupont,andY. Saeys, “Robust biomarker identification for cancer diagnosis with ensemble feature selection methods,” Bioinformatics, vol. 26, no. 3, pp. 392–398, 2010.

[2] D. Aha and D. Kibler, “Instance-based learning algorithms,” Mach. Learn., vol. 6, no. 1, pp. 37–66, 1991.

[3] 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.

[4] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays,” Proc. Nat. Acad. Sci., vol. 96, no. 12, pp. 6745–6750,1999.

[5] F. Alonso-Atienza, J. L. Rojo-Alvare, A. Rosado- Mu~noz, 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.