RESEARCH ARTICLE OPEN ACCESS

A Feature Selection Approach for Enhancing the Cardiotocography Classification Performance

Seema A Dongare[1], Vinay Kumar Ande [2], Ravi Kumar Tirandasu[3]
[1,2,3]Computer Engineering, Sanjivani College of Engineering, and Kopargaon

Abstract:

Data Mining is an interdisciplinary field of study for knowledge discovery from the large data sets. Collection of data, preprocessing, applying intelligent data mining techniques and interpretation are variou stages of data mining process. Out of which preprocessing is the critical part in data mining process. Noise, mislabelled data, imbalanced class, feature selection are some of the issues involved in preprocessing. In this research feature selection methodology has been considered for improving classification result for predicting Cardiotocography class. We have considered Correlation based Feature Selection(CFS), Symmetrical Uncertainty, ReliefF, Information Gain, Chi- Square feature selection methods and four different types of classifiers namely Jrip (Rule based), J48 (Tree based), NB (Bayes Learner), KNN (Lazy Learner). Proposed method has recorded some better result than considering all the features.

I. INTRODUCTION

In Today's information era huge amount of data is available which is of no use unless it is transformed into useful state. Data Mining is required to automatically analyze this huge amount of data to discover knowledge from it. We would be able to use this information in many applications such as Healthcare System, Fraud Detection, Market Analysis, Production Control and Science Exploration [1].

The term Knowledge Discovery from data (KDD) refers to the process of knowledge discovery in databases. The process of KDD is comprised of many steps, such as

- Data cleaning to remove noise and inconsistent data.
- Data integration to integrate data from multiple sources.
- Data selection to retrieve relevant data from the databases.
- Data transformation to transform data into suitable form.
- Data mining to apply intelligent techniques to discover patterns, pattern evaluation and knowledge representation to evaluate and visualize patterns.

Data mining has proven great benefits in healthcare domain. It uses data and analytics to identify the best practices to improve health condition of human beings. Processes are developed that make sure patients receive appropriate care at the right place and at the right time. Vast research has been done in this area to improve health of human beings.

Researchers have proposed a novel M-Cluster feature selection (Mcfs) based on Symmetrical Uncertainty (SU) Attribute Evaluator for improving the classification accuracy of medical datasets. Resultant feature subset has been tested on Dermatology and Breast Cancer medical datasets[2]. Data mining approach has been proposed by researchers for the classification of lung cancer subtypes[3]. Analysis of effectiveness of data mining techniques in healthcare system, various approaches, tools and its effect in healthcare has been presented by researchers [4].

Data preprocessing is an important step in data mining which addresses various issues as noise removal, class imbalancing and high dimensionality. Applying data mining techniques on high dimensional data is time consuming and is infeasible. So the Data reduction techniques like Data cube aggregation, Attribute subset selection, Dimensionality reduction, Numerosity reduction, Data discretization and Concept hierarchy generations can be used to deal with this issue. Data reduction must be applied in such a way that any valuable data will not be lost.

The performance of classification not only depends on classification technique being applied on processed data, but also on set of features present in the dataset. Dataset may include irrelevant and redundant features. If all features are selected for classification, its complexity will increase and the accuracy will reduce. So it is required to select subset of most relevant and non-redundant features which will maximize accuracy of classification and will minimize space and time requirement. This can be achieved using Feature Selection techniques.

There are three techniques available in the literature to achieve the best subset of features as Filter, Wrapper, and Hybrid. Using feature selection, best subset of features can be achieved, the memory required to perform data analysis is decreased, the speed of classification can be increased, and the most important of all is accuracy of classification model can be enhanced [5].

Filter method uses ranking technique for selection of relevant features which is independent of classifier. Features are ordered by score computed using suitable ranking criteria and threshold is set to discard irrelevant features. While in wrapper technique, a learning algorithm is used as a subroutine for evaluating the attributes importance in prediction over validation dataset. Embedded technique combines both filter and wrapper. Speed of computation as being an important parameter in classification, filter method is comparatively proven to be beneficial as being simple and provides good success ratio in classification. In this research work, filter methods are considered for comparing the feature subsets formed by proposed technique. Our work is tested over the cardiotocography data set.

Table 1 represents the list of filter based feature selection methods considered with their functional view.

Table 1. List of Feature Selection Methods

Name	**Functional View**
Symmetrical Uncertainty (SU)	Ranker
ReliefF (Rel)	Ranker
Information Gain(IG)	Ranker
Chi- Square (Chi)	Ranker

II. RELATED WORK

In data preprocessing class imbalance problem become greatest issue in data mining which occurs when class distribution among samples is not uniform. The most of classification algorithms biased towards the majority class if imbalanced problem is existed in the training dataset. This can be addressed using the oversampling technique called SMOTE [6]. Few researchers used the concept of SMOTE in the medical field to boost the classification performance by ensemble methods [7].

Many researchers have applied feature selection concept to overcome the problem of high dimensionality in dataset. Ensemble learning on two feature selection techniques as feature subset selection and feature rankers has been proposed for CTG data classification [8].

Feature selection based on correlation coefficient and Symmetrical Uncertainty is proposed by the researchers and applied over various dataset belongs to diverse areas [9].Symmetrical Uncertainty based feature selection method have been applied over some of the medical dataset to derive the best features before applying classification algorithms [10]. Researchers have proposed a novel method to address imbalanced data sets and small sample size. The issue of class distribution is handled by adding virtual samples generated using windowed regression oversampling (WRO) method [11].

Subset selection algorithm has been proposed by researchers which considers selected as well as remaining features which are relevant with the label. Their research work is intended for supervised classification which aims at selecting feature that can best predicts the class label [12].

A novel filter based approach for feature selection that sorts out the features based on a score and then performance of four different data mining classification algorithms on the resulting data has been analyzed by few researchers to investigate the

problem of efficient feature selection for classification on high dimensional datasets [13].

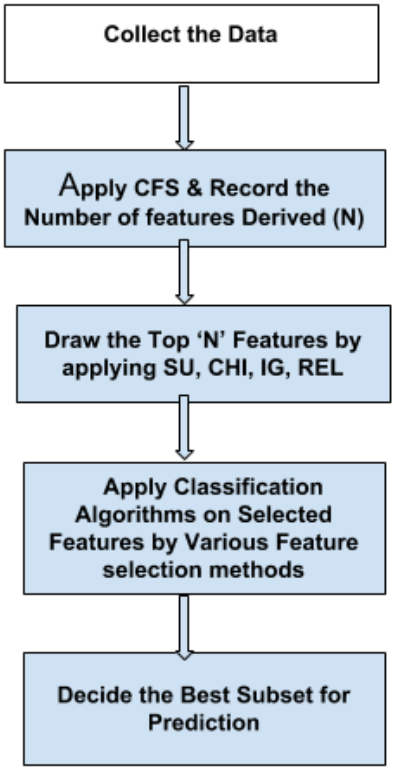
Researchers have proposed a fast clustering-based feature selection algorithm (FAST) which works in two steps. In the first step, grouping of features into clusters by using graph theoretic clustering methods is done. while in the second step, from each cluster the most representative feature that is strongly related to target class is selected to form a subset of features [14].

Gathering distributed data from multiple locations will not be economical or legal, so to deal with this distributed data, researchers have proposed a distributed approach for partitioned data using two techniques: horizontal (i.e. by samples) and vertical (i.e. by features). Then merging process using the theoretical complexity of these feature subsets has been proposed and analyzed [15].

Few researchers have proposed, an advanced novel heuristic algorithm, the hybrid genetic algorithm with neural networks (HGA-NN), which is used to identify an optimum feature subset and to increase the classification accuracy and scalability in credit risk assessment [16]

III. PROPOSED METHODOLOGY

The intention of proposed framework is to minimize the feature set for maximizing the accuracy of classification. From the data set consisting of 'F' features, if it is a required to select most relevant and non-redundant 'N' features, in such situation total C(F, N) number of subsets can be generated. Analyzing those many subsets in case of the high dimensional dataset is tedious and complex. In order to address this issue, filter based ranking techniques can be applied which will assign the rank to each feature and then most popular 'N' features can be considered for analysis. In this proposed methodology we consider CFS(Correlation based Feature Selection) for deriving the best features and further it is analyzed with some of the filter based ranking approaches like SU, IG, Chi, Rel. Proposed methodology works as per below flowchart.



Flowchart 1: Proposed Methodology

In order to analyze the performance of proposed framework, Cardiotocography data set is collected from UCI machine learning repository. The initial dataset has 2126 records, 23 features, and one Class label (Fatal State). CTGs are classified with respect to fatal state as (ONE, TWO, THREE). The dataset description is given in Table 2.

Table 2. Dataset Description

Title	Count
Total #Records	2126
Total #Features	23
Total #Classes	03

In second step, CFS is applied over the dataset and derived the best subset of features. Then recorded size of the subset (N). After this, filter based ranking approaches applied and considered top 'N' features for the analysis. Finally.classification techniques such as Jrip (Rule based), J48 (Tree based), NB (Bayes Learner),

KNN (Lazy Learner) are employed with the derived features and best subset is decided for the better classification. This experiment is carried out using WEKA machine learning tool with its default settings.

IV. RESULTS AND DISCUSSIONS

In this section obtained results with possible discussion is articulated. Performance of feature set selected using techniques as CFS and ranking based feature selection techniques CHI, IG, Rel, SU over all features is presented with brief discussion.

For the analysis of proposed methodology, Firstly CFS is applied over dataset which produces the best subset of 'N' features. Then ranking based feature selection techniques such as CHI, IG, Rel and SU are employed over the dataset. Top 'N' features are considered for further analysis. Best subset of 'N' features derived by CFS and topmost 'N' features with their ID's extracted by CHI, IG, Rel, SU are listed in Table 3.

Table 3: Features Derived by Various feature selection Methods

Feature selection Method	Selected Feature id's
CfsSubsetEval (CFS)	6,7,8,10,17,22
ChiSquaredAttributeEval (CHI)	22,10,18,9,8,17
Information Gain(IG)	22,9,10,8,18,20
ReliefFAttributeEval (Rel)	22,8,10,2,4,1
Symmetrical UncertAttributeEval (SU)	22,10,8,7,9,18

Further, feature set given by CFS and features selected by filter based methods are used for classification to decide the best feature set. Feature selection methods along with classification accuracies are represented in Table 4.

Table 4: Classification Result for Features Set Derived by Various Feature selection Methods

	Jrip	J48	NB	KNN
CFS	98.63	98.63	90.12	97.55
CHI	98.91	**98.73**	**92.09**	**98.07**
IG	98.77	**98.77**	91.48	97.78
REL	98.68	**98.77**	90.96	97.46
SU	**98.96**	**98.77**	90.35	97.83
ALL	98.73	98.58	86.78	96.89

The classification accuracy achieved with feature selection is better than that with all features. From the experimental analysis it can be observed that features selected using SU gives best accuracy with Jrip, J48. While features selected using CHI gives best result with J48, NB and KNN. Feature set derived using IG and Rel give almost equivalent and highest accuracy for tree based classifier. The comparison analysis of results are given in Figure 1.

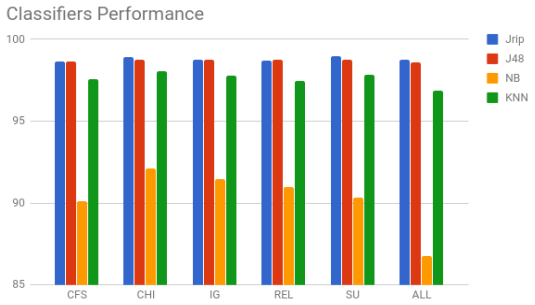


Figure 1: Classifier Performance

V. CONCLUSION

In this study, dimensionality reduction issue has been addressed. The proposed method is analyzed using Cardiotocography dataset. It has been observed that initial dataset has many features which are irrelevant and redundant. To deal with this problem, Initially, CFS features selection technique is applied to derive 'N' number of features, then ranking based filter techniques are applied and 'N' top most features are selected. All these feature sets are analyzed using JRip, J48, NB, KNN classifiers to select best feature set among all. Concluded results indicates that one of the subset, in some cases more than one subset giving improved results than all features of the dataset.

With this, we conclude that instead of selecting all features for classification, feature selection technique can be used for better classification accuracy.

REFERENCES

1. *Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.*
2. *Potharaju, S. P., &Sreedevi, M. (2017). A Novel M-Cluster of Feature Selection Approach Based on Symmetrical Uncertainty for Increasing Classification Accuracy of Medical Datasets. Journal of Engineering Science & Technology Review, 10(6).*
3. *Dass, M. V., Rasheed, M. A., & Ali, M. M. (2014, January). Classification of lung cancer subtypes by data mining technique. In Control, Instrumentation, Energy and Communication (CIEC), 2014 International Conference on (pp. 558-562). IEEE.*
4. *Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. Journal of healthcare information management, 19(2), 65.*
5. *Chandrashekar, G., &Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.*
6. *Chawla, N. V., Bowyer, K. W., Hall, L. O., &Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.*
7. *Potharaju, S. P., &Sreedevi, M. (2016). Ensembled Rule Based Classification Algorithms for predicting Imbalanced Kidney Disease Data. Journal of Engineering Science and Technology Review, 9(5), 201-207.*
8. *Silwattananusarn, T., Kanarkard, W., &Tuamsuk, K. (2016). Enhanced classification accuracy for cardiotocogram data with ensemble feature selection and classifier ensemble. Journal of Computer and Communications, 4(04), 20.*
9. *Potharaju, S. P., &Sreedevi, M. (2017). A Novel Clustering Based Candidate Feature Selection Framework Using Correlation Coefficient for Improving Classification Performance. Journal of Engineering Science and Technology Review, 10(6), 38-43.*
10. *Novaković, J. (2016). Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav Journal of Operations Research, 21(1).*
11. *Hu, Y., Guo, D., Fan, Z., Dong, C., Huang, Q., Xie, S., ...&Xie, Q. (2015). An improved algorithm for imbalanced data and small sample size classification. Journal of Data Analysis and Information Processing, 3(03), 27*
12. *Liu, Y., Tang, F., &Zeng, Z. (2015). Feature selection based on dependency margin. IEEE Transactions on Cybernetics, 45(6), 1209-1221.*
13. *Singh, B., Kushwaha, N., &Vyas, O. P. (2014). A feature subset selection technique for high dimensional data using symmetric uncertainty. Journal of Data Analysis and Information Processing, 2(04), 95.*
14. *Song, Q., Ni, J., & Wang, G. (2013). A fast clustering-based feature subset selection algorithm for high-dimensional data. IEEE transactions on knowledge and data engineering, 25(1), 1-14.*
15. *Morán-Fernández, L., Bolón-Canedo, V., & Alonso-Betanzos, A. (2017). Centralized vs. distributed feature selection methods based on data complexity measures. Knowledge-Based Systems, 117, 27-45.*
16. *Oreski, S., &Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert systems with applications, 41(4), 2052-2064.*