Adaptive Personalized Clinical Decision Support System Using Effective Data Mining Algorithms

Anima.P
M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India.

B. Senthil Kumar
Assistant Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India

Abstract – In the current health care unit, numerous and huge sized patient electronic records are collected and maintained. Data mining is the popular approach for handling such huge medical data. In this paper, an improved clinical decision support system with data mining techniques has developed. The personal health decision recommendation framework is designed and used to adapt different types of clinical datasets named as A-CDSS (adaptive clinical decision support system). The proposed framework has solved many research issues such as feature selection, classification issues like class inequality and accuracy problems and finally the decision selection based on the diagnosed result. The proposed system experimented with the huge sized three types of disease and patient electronic records which should be processed effectively. For this, the proposed system creates fast similarity and dissimilarity assessment for Personalized Clinical Decisions System (PCDS) framework with different algorithms and techniques such as Iterative-score-based Support Vector Machine (ISSVM) and Feature Selection using Fast similarity-Dissimilarity assessment (FSDA). The experiment result shows the proposed systems effectiveness in terms of accuracy.

Index Terms – Data mining, clinical decision support system, feature selection, health care data mining.

1. INTRODUCTION

Medical data processing is hard to deploy due to its dynamic and uncertain data formats. The clinical decision making from such data is very challenging for the physician and medical experts [1]. Data mining and decision support systems are the best method to overcome such issues. With the help of data mining, an effective clinical decision support system is developed with various research objectives [2]. Design and development of efficient feature selection and classification techniques for Clinical Decision Support System (CDSS) is identified as a critical research problem. Hence, a Clinical Decision Support System is proposed with multiple intension such as, a new feature selection approach and compare the performance of existing and the proposed feature selection algorithms on clinical datasets, to determine which feature selection technique and classification algorithm is best suited for clinical datasets under consideration, to design a SVM [3] based Classifier for different clinical disease diagnosis [4] and to predict the severity of disease in patients presenting with score value. And finally, to propose an appropriate decision based on the score.

From the detailed analysis of clinical decision support system in [5], an effective and optimal CDSS is proposed. However, there are several techniques and methods [6][7][8] have proposed to achieve the CDSS, its limited for certain medical data and yielded less accuracy. The summary of the existing clinical decision support techniques with classification problem have reviewed in [9]. And from the paper, the problem of existing algorithms have detected. In some existing studies [10], the authors concentrated on finding diabetic risks using data mining. From the above analysis, the proposed system introduced a new solution against different data mining issues related to the clinical data. The process and performance of proposed CDSS is discussed in this paper for various clinical datasets.

2. CLINICAL DECISION DISCOVERY ENGINE

Data mining is the popular area in the field of clinical support system. It uses machine learning, statistical and visualization techniques to discover and present knowledge in a form which is easily comprehensible to the users. It requires organization through large amounts of data and selecting appropriated and relevant information from the huge dataset. It is increasingly used in the sciences to extract information from the enormous data sets generated by modern trialing. In this paper, the use of knowledge discovery phase in clinical decision support system is explored. This will help for accurate Forecasting, or predictive modeling, which provides predictions of future events and may be transparent and readable in rule based systems. The concept of proposed system is, exploring the technique of extracting previously unknown information with the widest relevance from databases, in order to use it in the decision-making process.

A novel Clinical decision Discovery Engine is proposed in the paper, which optimizes clinical recommendations by identifying the features in the patient record that differentiate the individuals who receive a certain clinical decision. The approach utilizes the available contextual information about patients and learns from the large quantities of observational clinical data to inform clinical recommendations and make better decisions by learning from similar patients records. The proposed system has the following contributions.

- An adaptive A-CDSS approach for discovering the most relevant information from the patient health record (PHR) is proposed. This method distinguishes between patients that should receive one particular clinical decision and the patients who should be given another. For instance, different types of disease datasets are used.
- With the use of the past records in the PHR and external knowledge from the medical literature, the approach discovers the optimal personalized clinical decision based on the discovered relevant information's.
- For effective clinical decision support system, FSDA based feature selection is proposed.

Improved Clinical Decision Recommendation Algorithm is proposed. This is based on the advanced classifier of Support vector machines with iterative score based. The proposed classifier is named as ISSVM.

- The proposed system utilizes the score of the class and finds appropriate decision for the score value this type of classification overcomes the problem of over fitting and classification accuracy problem.
- The A-CDSS is applied in three medical data sources:

1. Personalized treatment recommendations for liver, heart and diabetic patients and

2. Diagnosis of liver, heart and diabetic disorders.

- A-CDSS is used to discover the most appropriate features to make a distinct clinical decision and then uses this knowledge to build a clinical decision recommendation system.
- The system performs the evaluation of A-CDSS in the context of multiple disease diagnosis and treatment and show that it consistently and significantly outperforms state-of-the-art machine learning algorithms.

2.1. Proposed Clinical Decision Support System (CDSS):

The proposed system developed a CDSS for multiple clinical data sources with performance improvements. This consists of an iterative sequence of the common data mining process shown in Figure 1.0:

The data cleaning phase refers to the preprocessing of data in order to remove or reduce noise. This will performed in the proposed system by applying smoothing techniques and this treats the missing values also. The missing value will be replaced with the most commonly occurring value for that attribute or with the most probable value based on statistics.

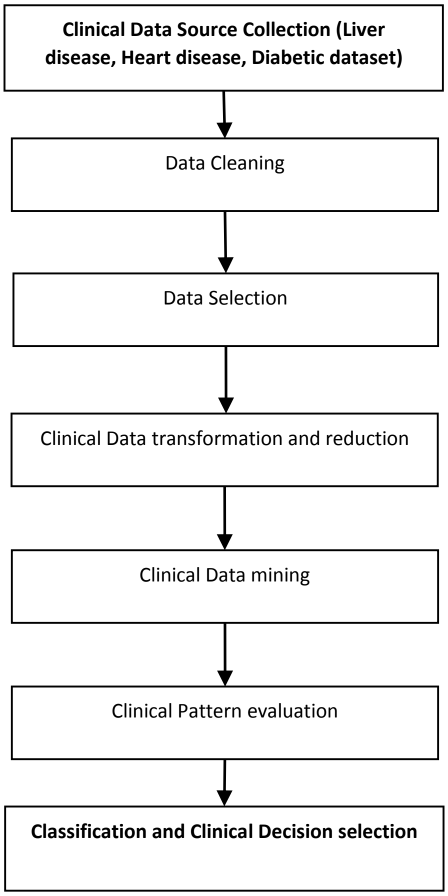

Clinical Data Source Collection (Liver disease, Heart disease, Diabetic dataset)
Data Cleaning
Data Selection
Clinical Data transformation and reduction
Clinical Data mining
Clinical Pattern evaluation
Classification and Clinical Decision selection

Figure 1.0 process involved in the proposed CDSS Data cleaning

Data Selection-

The proposed system experiments the CDSS with several clinical dataset, which may have many of the attributes will be redundant. Shannon entropy is calculated for pre-processing and data selection. This can be used to identify whether any two given attributes are statistically related. A database may also contain irrelevant attributes. Attribute subset selection can be used in these cases to find a reduced set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. Hence, the calculation, in the form of pre-processing and attribute subset selection, can be used to detect attributes that do not contribute to the classification or prediction task.

Clinical Data transformation and reduction-

The collected clinical data's are sometimes in different formats, which may be transformed by normalization, particularly when decision support system utilizes the open source datasets and this involving distance measurement are used in the learning step. Normalization involves scaling all values for a given attribute so that they fall within a small specified range, such as -1.0 to 1.0, or 0.0 to 1.0.

Clinical Data mining-

It is an essential process where intelligent methods are applied in order to extract data patterns to find the class and for classification. It performs functional tasks such as classification, association, characterization, cluster analysis and prediction of diseases.

Clinical Pattern evaluation –

Clinical patterns are unpredictable and uncertain, so this process is used to identify the truly appropriate patterns representing knowledge based on some similarity and dissimilarity measures.

Decision Selection –

Here appropriate decision selection and visualization of clinical results are performed to present the mined knowledge to the user.

1.2 Proposed Methodology:

1.2.1 Feature Selection using Fast similarity-Dissimilarity assessment (FSDA):

In the initial step of the proposal includes the effective feature selection, which can detect the class rapidly and accurately in the A-CDSS. FCBF is a procedure proposed for effective feature selection, which is based on attribute association analysis of with given classes. This involves two views:

1. One is the detection of similarity and dissimilarity calculation for every feature in the feature set.

2. The second process is identifying the relevant feature is redundant or not. Based on these two a Statlogs, the finalized features are selected. This selection will be performed after comparing with every feature. To define the similarity and dissimilarity a user-defined threshold *T* value, as the method used by the feature weighting algorithms. More specifically, suppose a data set *S* contains *N* features and a class *C*. Let $T_{i,c}$ denote the *T* value that measures the correlation between a feature F_i and the class C then a subset S' of relevant features can be decided by a threshold T value į, such that F_i S ,1 i N, $T_{i,c}$ į. The second process is more complicated because it may involve analysis of pair-wise correlations between all features, which results in a time complexity of $O(N^2)$ associated with the number of features N for most existing algorithms.

Algorithm

```
input: S(F_1, F_2, . . . , F_N, C) // a training data set
// a predefined threshold // an optimal subset
for i = 1 to N do begin
calculate $T_{i,c}$ for $F_i$;
if ($T_{i,c}$ ш )
append $F_i$ to $S_{list}$;
end;
order $S_{list}$ in descending $T_{i,c}$ value;
$F_p$ = getFirstElement($S_{list}$);
do begin
$F_q$ = getNextElement($S_{list}$; $F_p$);
if ($F_q$<> NULL)
do begin
$F'_q$ = $F_q$;
if ($T_{p,q}$ ш $T_{q,c}$)
remove$F_q$ from $S_{list}$;
$F_q$ = getNextElement($S_{list}$, $F'_q$);
else        $F_q$ = getNextElement($S_{list}$, $F_q$);
end until ($F_q$ == NULL);
$F_q$ = getNextElement($S_{list}$, $F_q$);
end until ($F_p$ == NULL);
Sbest = S list;
end;
```

The FSDA output will be applied into the A-CDSS. This improves the diagnosis accuracy and decision selection process among different types of datasets.

1.2.2 Classification Process:

In this research work, FSDA is used for Feature extraction and Iterative-score-based Support Vector Machine (ISSVM), a semi-supervised self learning method is used for disease diagnosis. Feature extraction changes the information in the high-dimensional space to a space of lower measurements. Feature extraction includes disentangling the measure of assets required to depict an expansive arrangement of information precisely. When performing examination of complex information one of the real issues comes from the quantity of factors included. Feature extraction is a general term for strategies for building blends of the factors to get around these issues while as yet portraying the information with adequate precision. Best outcomes are accomplished when a specialist develops an arrangement of utilization subordinate features. Feature extraction is implemented using the FSDA.

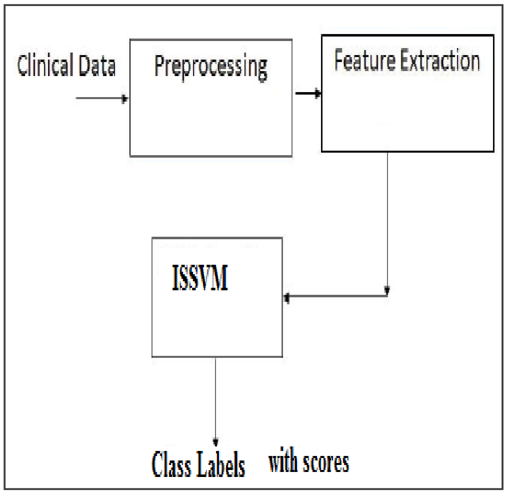


Figure 2.0 process of ISSVM classification

ISSVM have been successfully applied on three types of disease dataset. It is primarily applied a classifier that performs classification tasks by constructing an optimal score based hyper plane in a high dimensional space that segments members and non-members of a given class with the execution of similarity and dissimilarity learning. ISSVM is more effective for clinical data analysis for its efficiency in handling clinical datasets, its flexibility in choosing a similarity function and its ability to identify the class accurately. The overall process of ISSVM is defined in figure 2.0. ISSVM is a semi supervised machine learning algorithm based on statistical learning theory. The features that make ISSVM attractive for Pattern classification problems are class discrepancy issues, good generalization performance, Computational efficiency, and good performance in higher dimensions. For a k-class problem, it constructs *k* binary classifiers with i^{th} classifier trained to discriminate samples of i^{th} class from other classes. The proposed classifier has training and testing phases, which are described below.

3. IMPLEMENTATION AND RESULTS

Three clinical datasets of varying sizes and characteristics were obtained from UCI Machine Learning Repository used for the analysis. The details of the datasets are as follows:

3.1 Dataset Used:

A. Diabetes Dataset

The proposed system initially performs the experiment on Diabetes dataset [11]. The first dataset namely the UCI Diabetes dataset is used to discriminate healthy people from those with Diabetes disease, according to class attribute which is set to either 0 for healthy and 1 for Diabetes disease. This dataset contains 19 numeric valued attributes and 1 binary valued class variable and 114 patient records. The above details show the diabetes dataset attributes and total instances retrieved from the UCI repository.

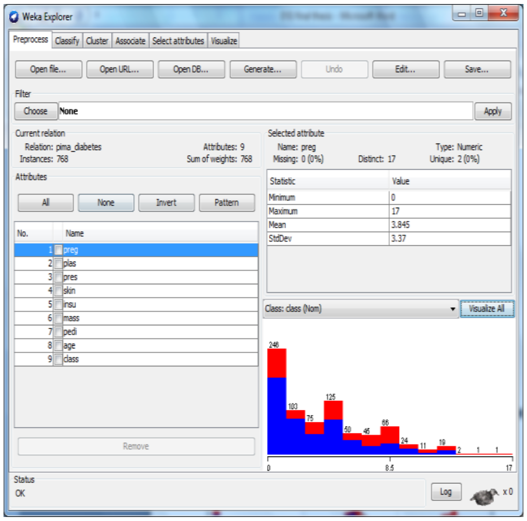
B. LIVER Dataset

LIVER' dataset [12]classification and decision support system task is to distinguish disease versus normal patterns from mass liver dataset. This is a two-class classification problem with continuous input variables. LIVER dataset was obtained by merging three mass-Statlogrometry datasets to obtain enough training and test data for a benchmark. The original features indicate the abundance of certain features in human having a given mass value. Based on those features one must separate liver patients from healthy patients. The order of the features and patterns were randomized. This dataset contains 345 patient records with 6 attributes. The dataset is divided into two datasets: Training Dataset with 345 samples Testing Dataset with 175 samples. The algorithms were trained using the Training Dataset and evaluated using the Testing Dataset.

C. STATLOG Heart Dataset

Statlog Heart dataset describes diagnosing of cardiac heart disease dataset [13]. Each patient is classified into two categories: normal and abnormal. The database of 267 Statlog has been processed to extract features that summarize the original heart datasets. This contains 270 instances and 13 attributes. The dataset is divided into two datasets: Training Dataset with 150 instances Testing Dataset with 120 instances. The algorithms were trained using the Training Dataset and evaluated using the Testing Dataset.

3.2 Experimental Results



Filter: None

Current relation
Relation: pima_diabetes Attributes: 9
Instances: 768 Sum of weights: 768

Selected attribute
Name: preg Type: Numeric
Missing: 0 (0%) Distinct: 17 Unique: 2 (0%)

Statistic	Value
Minimum	0
Maximum	17
Mean	3.845
StdDev	3.37

Attributes

No.		Name
1	□	preg
2	□	plas
3	□	pres
4	□	skin
5	□	insu
6	□	mass
7	□	pedi
8	□	age
9	□	class

Class: class (Nom)

Status
OK

Figure 3.0 the initial analysis of the liver dataset results in the Weka tool.

The proposed algorithms were run on the above three datasets in two software's such as Weka and C#.net. The performance of the feature selection algorithms were measured by using the three classifiers on the selected features. Initially, the same datasets are converted into the “arff” format to run on weka tool. Using the experiment, the results can be compared with the proposed system.

The proposed system shows that diagnosis and selection of CDSS by FSDA and ISSVM based classifiers can provide high degree accuracy in the liver, heart and diabetes clinical datasets. The accuracy of feature extraction methods with different ISSVM kernels for different clinical datasets is presented.

Table 1.0 Classification accuracy of different feature selection techniques on Clinical Dataset

Algorithm	**Liver**	**Diabetes**	**Heart**
Genetic Algorithm	96.23%	96.16%	92.05%
PCA	95.67%	95.36%	94.41%
Chi square	96.05%	95.26%	93.14%
FSDA	98.49%	98.11%	98.06%

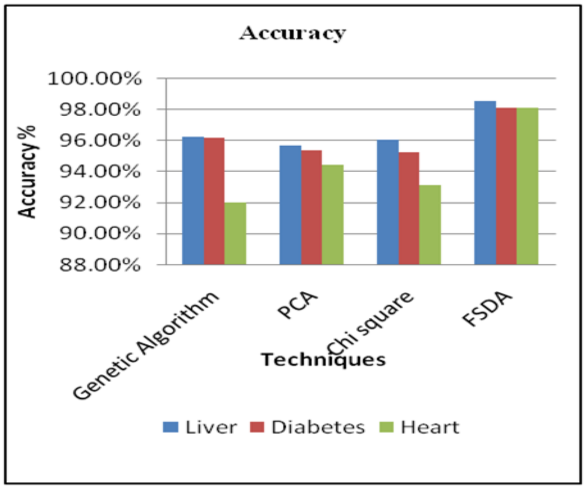


Figure 4.0overall comparison for all dataset

It is evident from the results that out of the existing Classifier, the proposed system performs exceptionally well for Diabetes, liver and heart dataset. It has nearly the same performance with all the feature selection algorithms. The FSDA-ISSVM combination has better performance in STATLOG Heart and LIVER datasets. The percentages of the predicted values are match with the expected value for the given data. The best system is that having the high Accuracy, High Precision and High Recall value. The performance of the proposed system is tested with the three dataset with multiple instances, from each instance the precision and recall values are gathered and that is plotted in the above figure 4.0.

4. CONCLUSION

The main highlight of the present research work is to design and develop An Adaptive Clinical Decision Support System for the liver, Heart and diabetes Diseases” to improve the Medical Diagnosis and Decision Management. A brief summary of this paper is given as follows The present study is initiated with an introductory note on Heart Diseases, liver Diseases, and Decision Support systems. The proposed system performed effective feature selection technique for effective data classification and decision selection. To achieve the higher accuracy, the system proposed an effective clinical decision support system, FSDA based feature selection. In this paper, an advanced classifier of Support vector machines with iterative score based schemes are adapted with the FSDA for better classification. The proposed classifier is named as ISSVM. The proposed system utilizes the score of the class and finds appropriate decision for the score value this type of classification overcomes the problem of over fitting and classification accuracy problem. The A-CDSS is applied in three medical data sources such as heart, liver and diabetes disease datasets from UCI repository. The system performed two types of experiments, one in Weka for existing algorithm performance analysis and proposed in C#.net software. A-CDSS is used to discover the most appropriate features to make a distinct clinical decision and then uses this knowledge to build a clinical decision recommendation system. The system performed the evaluation of A-CDSS in the context of multiple disease diagnosis and treatment and show that it consistently and significantly outperforms state-of-the-art machine learning algorithms. Finally the experiments show the proposed system gained better result.

REFERENCES

[1] Croskerry, Pat. "Individual variability in clinical decision making and diagnosis." *Diagnosis: Interpreting the Shadows. Oxford, UK: CRC Press, Taylor Francis Group* (2017).

[2] Musen, Mark A., Blackford Middleton, and Robert A. Greenes. "Clinical decision-support systems." *Biomedical informatics*. Springer London, 2014. 643-674.

[3] Kumari, V. Anuja, and R. Chitra. "Classification of diabetes disease using support vector machine." *International Journal of Engineering Research and Applications* 3.2 (2013): 1797-1801.

[4] Senthil Kumar, B & Gunavathi R., Dr. (2016). A Survey on Data Mining Approaches to Diabetes Disease Diagnosis and Prognosis. IJARCCE. 5. 463-467. 10.17148/IJARCCE.2016.512105.

[5] Kumar, B. Senthil. "Data Mining Methods and Techniques for Clinical Decision Support Systems." *Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org* 7, no. 8 (2017).

[6] Hermsen, E. D., VanSchooneveld, T. C., Sayles, H., & Rupp, M. E. (2012). Implementation of a clinical decision support system for

antimicrobial stewardship. Infection Control & Hospital Epidemiology, 33(4), 412-415.

[7] Lomotan, E. A., Hoeksema, L. J., Edmonds, D. E., Ramírez-Garnica, G., Shiffman, R. N., & Horwitz, L. I. (2012). Evaluating the use of a computerized clinical decision support system for asthma by pediatric pulmonologists. International journal of medical informatics, 81(3), 157-165.

[8] Kumar, D. Senthil, G. Sathyadevi, and S. Sivanesh. "Decision support system for medical diagnosis using data mining." International Journal of Computer Science Issues 8.3 (2011): 147-153.

[9] Kumar, B. Senthil, and R. Gunavathi. "Comparative and Analysis of Classification Problems." *Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org* 7.8 (2017).

[10] Senthil Kumar, B & Sreejith.R. (2016).”A Survey on Identification of Diabetes Risk Using Machine Learning Approaches”. IJIRCCE. 4. 33 - 335 . 10.15680/IJIRCCE.2016. 0408001.

[11] https://archive.ics.uci.edu/ml/datasets/diabetes

[12] https://archive.ics.uci.edu/ml/datasets/liver+disorders

[13] https://archive.ics.uci.edu/ml/datasets/spect+heart