Introduction to Feature Subset Selection Method

Hemal Patel[1] Mr. Lokesh Gagnani[2] Mrs. Mansi Parmar[3]
[1]M.E Scholar [2,3]Assistant Professor
[1,2,3]Department of Information Technology
[1,2,3]Kalol institute of technology – India

Abstract— Data Mining is a computational progression to ascertain patterns in hefty data sets. It has various important techniques and one of them is Classification which is receiving great attention recently in the database community. Classification technique can solve several problems in different fields like medicine, industry, business, science. PSO is based on social behaviour for optimization problem. Feature Selection (FS) is a solution that involves finding a subset of prominent features to improve predictive accuracy and to remove the redundant features. Rough Set Theory (RST) is a mathematical tool which deals with the uncertainty and vagueness of the decision systems.

Key words: Classification, Particle Swarm Optimization (PSO) Rough Sets, Feature Selection (FS)

I. INTRODUCTION

Data mining is the process of selecting, exploring and modelling large amounts of data in order to discover unknown patterns or relationships which provide a clear and useful result to the data analyst [1]. There are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on available data.

Data mining involves some of the following key steps:

1) Problem definition: The first step is to identify goals.
2) Data exploration: All data needs to be consolidated so that it can be treated consistently.
3) Data preparation: The purpose of this step is to clean and transform the data for more robust analysis.
4) Modelling: Based on the data and the desired outcomes, a data mining algorithm or combination of algorithms is selected for analysis. The specific algorithm is selected based on the particular objective to be achieved and the quality of the data to be analysed.
5) Evaluation and Deployment: Based on the results of the data mining algorithms, an analysis is conducted to determine key conclusions from the analysis and create a series of recommendations for consideration.

A. Techniques of Data Mining:

There are several major data mining techniques have been developing and using in data mining projects recently including association, classification, clustering, prediction, sequential patterns and decision tree.

1) Association:

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. That's is the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together.

2) Classification:

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we develop the software that can learn how to classify the data items into groups. For example, we can apply classification in application that "given all records of employees who left the company, predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups.

3) Clustering:

Clustering is a data mining technique that makes meaningful or useful cluster of objects which have similar characteristics using automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. To make the concept clearer, we can take book management in library as an example. In a library, there is a wide range of books in various topics available. The challenge is how to keep those books in a way that readers can take several books in a particular topic without hassle. By using clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in that topic, they would only have to go to that shelf instead of looking for entire library.

4) Prediction:

The prediction, as it names implied, is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. For instance, the prediction analysis technique can be used in sale to predict profit for the future if we consider sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

5) Sequential Patterns:

Sequential patterns analysis is one of data mining technique that seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period.

In sales, with historical transaction data, businesses can identify a set of items that customers buy together a different times in a year. Then businesses can use this information to recommend customers buy it with better deals based on their purchasing frequency in the past.

6) Decision trees:

Decision tree is one of the most used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision

tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it. For example, We use the following decision tree to determine whether or not to play tennis.

II. CLASSIFICATION

Classification involves predicting an outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, normally known as prediction attribute. The algorithm discovers the relationships between the attributes that would make it possible to predict the outcome. After that the algorithm is given a new data set called prediction set, which contains the same set of attributes, except for the prediction attribute is not yet known. The algorithm analyses the input and generates a prediction.

A. Classification Discovery Models[2]:

1) Decision Tree:

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modelling approaches used in statistics, data mining and machine learning.

Decision trees used in data mining are of two main types:

- Regression tree analysis is when the predicted outcome can be considered a real number (e.g. the price of a house,or a patient‟s length of stay in a hospital).
- Classification tree analysis is when the predicted outcome is the class to which the data belongs.

B. Neural Networks:

Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques.

A neural network consists of interconnected processing elements also called units, nodes, or neurons. The neurons within the network work together, in parallel, to produce an output function. Since the computation is performed by the collective neurons, a neural network can still produce the output function even if some of the individual neurons are malfunctioning (the network is robust and fault tolerant).

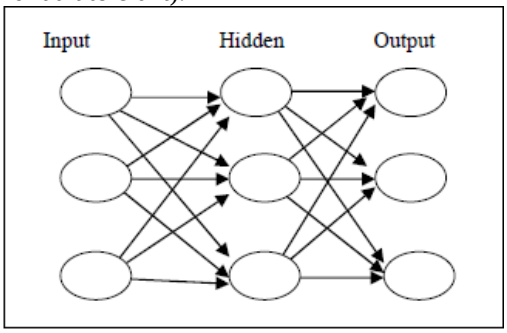

Fig. 1: Layer Of Neural Network

1) Genetic Programming:

Genetic programming (GP) has been vastly used in research in the past 10 years to solve data mining classification problems. The reason genetic programming is so widely used is the fact that prediction rules are very naturally represented in GP. Additionally, GP has proven to produce good results with global search problems like classification. GP consists of stochastic search algorithms based on abstractions of the processes of Darwinian evolution.

2) Fuzzy Sets:

Fuzzy sets form a key methodology for representing and processing uncertainty. Fuzzy sets constitute a powerful approach to deal not only with incomplete, noisy or imprecise data, but may also be helpful in developing uncertain models of the data that provide smarter and smoother performance than traditional systems.

- In Classification collected data is usually associated with a high level of noise. There are There are many reasons causing noise in these data, among which imperfection in the technologies that collected the data and the source of the data itself are two major reasons.

Dimensionality reduction is one of the most popular techniques to remove noisy (i.e. irrelevant) and redundant features.

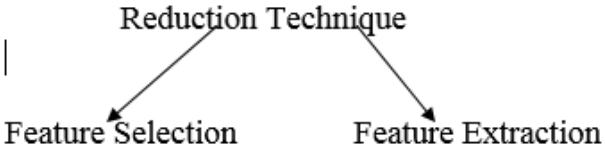

Fig. 2:

3) Feature Extraction:

Feature extraction approaches project features into a new feature space with lower dimensionality and the new constructed features are usually combinations of original features.

Example: Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Canonical Correlation Analysis (CCA).

III. FEATURE SELECTION

It aim is to select a small subset of features that minimize redundancy and maximize relevance to the target such as the class labels in classification.

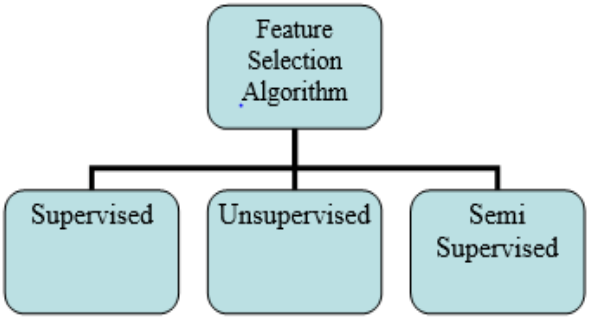

Fig. 3:

A feature selection method consists of four basic steps namely, subset generation, subset evaluation, stopping criterion, and result validation.

1) A candidate feature subset will be chosen based on a given search strategy, which is sent,
2) To be evaluated according to certain evaluation criterion.

3) The subset that best fits the evaluation criterion will be chosen from all the candidates that have been evaluated after the stopping criterion are met.
4) The chosen subset will be validated using domain knowledge or a validation set.

feature selection selects a subset of features from the original feature set without any transformation, and maintains the physical meanings of the original features.

feature selection for classification attempts to select the minimally sized subset of features according to the following criteria,

- The classification accuracy does not significantly decrease.
- The resulting class distribution, given only the values for the selected features, is as close as possible to the original class distribution, given all features.

IV. SUBSET SELECTION

Subset selection evaluates a subset of features as a group for suitability. Many popular search approaches use greedy hill climbing, which iteratively evaluates a candidate subset of features, then modifies the subset and evaluates if the new subset is an improvement over the old.

Alternative search-based techniques are based on targeted projection pursuit which finds low-dimensional projections of the data that score highly: the features that have the largest projections in the lower-dimensional space are then selected.

A. Search approaches include:

- Exhaustive
- Best first
- Simulated annealing
- Genetic algorithm
- Greedy forward selection
- Greedy backward elimination
- Particle swarm optimization
- Targeted projection pursuit
- Scatter Search
- Variable Neighborhood Search

V. PARTICLE SWARM OPTIMIZATION (PSO)

Particle Swarm Optimization (PSO) is a conventional and semi-robotic algorithm. It is based on the social behaviour associated with bird's flocking for optimization problem. A social behaviour pattern of organisms that live and interact within large groups is the inspiration for PSO. The PSO is easier to lay into operation than Genetic Algorithm. It is for the motivation that PSO doesn't have mutation or crossover operators and movement of particles is effected by using velocity function[3].

PSO, Particle Swarm consists of 'n' particles. The position of each particle stands for potential solution in D-dimensional space. Individuals, potential solutions, flow through hyper dimensional search space. The experience or acquired knowledge about its neighbours influences the changes in a particle within the swarm. The PSO algorithm involves of just three steps, which are being replicated until stopping condition, they are as follows[4].

1) Evaluate the fitness of each particle.
2) Update individual and global best functions.
3) Update velocity and position of each particle.

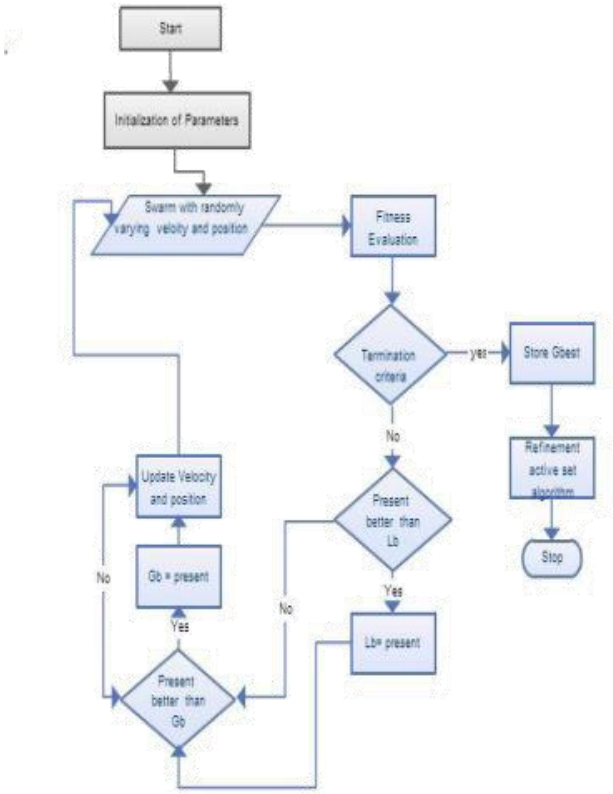

Fig. 4: Working of PSO

VI. ROUGH SET THEORY

Rough set theory can be regarded as a new mathematical tool for imperfect data analysis. The theory has found applications in many domains, such as decision support, engineering, environment, banking, medicine and others.

A. Advantages:

- It provides efficient methods, algorithms and tools for finding hidden patterns in data.
- It allows to evaluate the significance of data.
- It allows to generate in automatic way the sets of decision rules from data.
- It is easy to understand.
- It offers straightforward interpretation of obtained results.

VII. LITERATURE REVIEW

Sr No	Method Name	Description	Advantage
1	SPSO-QR[5]	It start with an empty set and it adds one at a time , in turn.	The dependency of subset is calculated based on dependency & decision attribute and best particle is chosen
2	SPSO-RR[5]	It start by selecting random values for each particle &	To avoid calculation of discern ability functions which can be computationally expensive without

		velocity.	optimizations.
3	PSO[6]	It is based based on the use of multiple sub-swarms instead of one (standard) swarm.	It increase overall performance of network.
4	HGAPSO[7]	It is obtained through integrating standard velocity & update rules of PSO with selection, crossover & mutation from the GA.	It does not need to set the number of desired features a priori.

Table 1:

VIII. CONCLUSIONS

In this paper , we introduce the Feature Selection and Subset Selection different method with comparison of other methods. It is used for medical data set.PSO method is used at beginning it increase performance of network. To predicate accuracy level hybridize PSO method with Rough Set Theory.

ACKNOWLEDGMENT

I am extremely obliged to my guide Mansi Parmar and Mr. Lokesh Gagnani devoid of them guidance the work would not have happened and They supported me to solve my difficulties arise and give Valuable suggestions. I would like to pay my sincere gratitude for them endless motivation and support in progess and success of this survey work.

REFERENCES

[1] A.Shameem Fatima , D.manimegalai and Nisar Hundewale , "A Review of Data Mining Classification Techniques Applied for Diagnosis and Prognosis of the Arbovirus-Dengue":Ijcsi -Volume 8, Issue 8, No.3 November 2011.
[2] Ritika MTech Student , "Research on Data Mining Classification":Ijarcsse - Volume 4, Issue 4, April 2014.
[3] Sivagowry.S , Dr. Durairaj.M, "PSO - An Intellectual Technique for Feature Reduction on Heart Malady Anticipation Data":Ijarcsse - Volume 4, Issue 9, September 2014 .
[4] Durairaj. M , Sivagowry. S, "Feature Diminution by Using Particle Swarm Optimization for Envisaging the Heart Syndrome": Ijitcs- January 2015 .
[5] H.Hannah Inbarani,Ahmad Taher Azar and G. Jothi "Supervised hybrid feature selection based on PSOand rough sets for medical diagnosis": Elsevier- 2014 .
[6] Sivagowry.S , Dr. Durairaj.M, "PSO - An Intellectual Technique for Feature Reduction on Heart Malady Anticipation Data":Ijarcsse - Volume 4, Issue 9, September 2014 .
[7] Pedram Ghamisi, Student Member, IEEE, and Jon Atli Benediktsson, Fellow, IEEE, "Feature Selection Based on Hybridization of Genetic Algorithm and Particle Swarm Optimization":IEEE - Volume 12, No 2 February 2015.

