Performance Analysis of PSO-KStar Classifier over Liver Diseases

P. Thangaraju1, R. Mehala2

Abstract— **The objective of this paper is to analyze the data of liver diseases using particle swarm optimization algorithm (PSO) with KStar Classification, in two aspects for classifying the existence of disease or not. A prototype was proposed to find the chances of occurrence of liver diseases on the basis of input variables by building an intelligent system based on feature selection. With the proof concept, the proposed algorithm enhanced the performance of accuracy when compared to existing classification algorithms. The results of the algorithm were analyzed and presented in this paper.**

Index Terms— **PSO, DM, Classification, MDM, Kstar.**

I. INTRODUCTION

Data Mining (DM) is the process of discovering knowledge or facts that are hidden from large data stores. Mining useful information helps users in taking right decision at right time [7]. Data mining algorithms are developed to be incorporated in various phases of DM such as preprocessing, feature selection, classification, clustering and visualization.

Clinical databases contain large quantities of information about patients and their medical conditions. Relationships and patterns within this data could provide new medical knowledge [8]. Data mining methods assist physicians in numerous ways right from the interpretation of complex diagnostic tests, merging information from multiple sources and providing support for differential diagnosis and providing patient-specific prognosis [9]. The largest and perhaps the most resistant of all the organs in the body, the liver is also one of the most mysterious. It is in fact responsible for over 500 functions including regulating controlling cholesterol levels, sex hormones, and vitamin and mineral supplies, warding off viruses and disposing of toxic material from the blood. It is the only organ that has the ability to rejuvenate [1].

Classification algorithms are procedures that select a hypothesis from a set of alternatives that best fits a set of observations. It maps sets to hypotheses from training so as to minimize the objective function. Classification algorithms play a major role in medical data mining for diagnosing the presence of disease as it replaces the intervention of physicians by improving the prediction accuracy of diseases

even in critical situations[10]. Many classification algorithms are introduced to classify and analyze the medical dataset. In this paper, a novel,PSO algorithm was implemented to classify the disorders that could possibly causes problems to liver.

Manuscript received June, 2015.

P. Thangaraju1*, Associate Professor, Department of Computer Application, Bishop Heber College (Autonomous), Tiruchirapalli, India1*

R. Mehala2*, M. Phil., Research Scholar, Dept., of Computer Science, Bishop Heber College (Autonomous), Tiruchirapalli, India2*

II. RELATED WORK

Murugan et al. [1], analyzed the liver cancer DNA sequence data using the generalization of Kimura Models and Markov Chain. They focused at the analysis of biological modules levels than the individual genes. Their approach produced results that were biologically interpreted and statistically robust. They insisted upon the use of biological knowledge for developing analytic techniques. Their experiment revealed that the percentage of accuracy was approximately same for all states.

Chen et al. [2], established a survival prediction model for liver cancer using ANN, Classification and Regression Trees. They tested the proposed model with three conditions, clinical stage alone, significant variables alone, and both significant and non-significant variables. They predicted the most influencing attributes of five year survival of patients. They found that the ANN model was found to be more significant and accurate compared to the other two models.

Rajeswari et al. [3], compared the classification accuracy of liver disorder dataset using naïve bayes, KStar and FTree Algorithms. The classification model was built and processed using WekaTool. They found that the classification accuracy of the FT algorithm was better than the KStar and FTree algorithms with 97.10%.

Aneeshkumar et al. [4], focused on the estimation on the surveillance of liver disorder using naïve bayes and C4.5 classification algorithms. They tested the surveillance of the patient with 15 influencing attributes. They found that the overall performance of C4.5 algorithm was better than the naïve bayes classifier.

Sug et al. [5], suggested an over-sampling method for minor classes to compensate the effectiveness with inefficient data using C4.5 and CART algorithms. When they increased the number of instances of minor classes by duplication they were able to achieve better classification results with the C4.5 and CART algorithms

Anthonyaet al.[11], implemented an exhaustive search method (also called enumerative search method) works by considering all possible band combinations by way of calculating their separability indices. Although this search method guarantees the optimality of solution, it poses the problem of being computationally prohibitive. For a dataset with d features (i.e. bands), 2d – 1 combinations are possible.

This method is practicable if the number of bands is less than 10. The use of 10 or more bands would be costly in terms of computational speed. However are of the opinion that advancements in computer technology should eventually render exhaustive search an operational reality. This, including the fact that the datasets considered in this research had less than ten bands, influenced the author's decision to consider this method.

Hall[12], used best first search is an Artificial Intelligence search strategy that allows backtracking along the search path. Best first moves through the search space by making local changes to the current feature subset. However, if the path being explored begins to look less promising, the best first search can back-track to a more promising previous subset and continue the search from there. Given enough time, a best first search will explore the entire search space, so it is common to use a stopping criterion. Normally this involves limiting the number of fully expanded subsets that result in no improvement.

K.Revathiet al. [13], employed probabilistic Search: LVF: Las Vegas Filter algorithm adopts the inconsistency rate as the evaluation measure. It generates feature subsets randomly with equal probability, and once a consistent feature subset is obtained that satisfies the threshold inconsistency rate. LVF is fast in reducing the number of features in the early stages and can produce optimal solutions.

Aziz et al. [14],instigated genetic Algorithms (GA) are search algorithms inspired by evolution and natural selection, and they can be used to solve different and diverse types of problems. The algorithm starts with a group of individuals (chromosomes) called a population. Each chromosome is composed of a sequence of genes that would be bits, characters, or numbers. Reproduction is achieved using crossover (2 parents are used to produce 1 or children) and mutation (alteration of a gene or more). Each chromosome is evaluated using a fitness function, which defines which chromosomes are highly-fitted in the environment. The process is iterated for multiple times for a number of generations until optimal solution is reached. The reached solution could be a single individual or a group of individuals obtained by repeating the GA process for many runs.

Jantawanet al. [15] implemented greedy search is a discrete version of the gradient descent (ascent) algorithm, implements a local search of each repetition. The algorithm starts with an initial network and determines a nearest neighbor graph that improves the network by including, eliminating or inverting an arc in the graph. The process is repeated until there is no neighbor that improves the current solution.

Kumariet al. [16] used forward selection: This method starts with no variables. Add the variables one by one, at each step adding the feature that has the minimum error. Repeat the above step until any further addition does not signify any decrease in error. Backward selection: This method starts with all variables. It then removes the variables one by one, at each step removing the feature that has the highest error. Repeats the above step until any further removal increases the error significantly

III. METHODOLOGY

The proposed method is implemented using an evolutionary algorithm called Particle Swarm Optimization (PSO) with KStar classification algorithm for classifying liver cancer. PSO is a heuristic global optimization method and also an optimization algorithm, which is based on swarm intelligence [6]. It is originated from the research on the bird and fish flock movement behavior. The concepts of PSO can be extensively applied in any domain as its results could give better solutions ever. The algorithm is widely used and quickly developed for its easy implementation by tuning only few particles.

In PSO a group of random particles are initialized for searching optimum solutions by updating the generations. In every epoch, each particle is updated with two best values called pbest and gbest. pbest is a best solution(fitness) that has achieved so far. Gbest is the best solution tracked and obtained by the PSO so far through any particle in the population. When a particle generates population with its topological neighbors, a local best value called lbest is generated.

Once when the two best values are found, the particle updates its velocity and positions with following equation (a) and (b).

v[] = v[] + c1 * rand() * (pbest[] - present[]) + c2 * rand() * (gbest[] - present[]) (a)

present[] = persent[] + v[] (b)

v[] is the particle velocity, persent[] is the current particle (solution). pbest[] and gbest[] are defined as stated before. rand () is a random number between (0,1). c1, c2 are learning factors. usually c1 = c2 = 2.

The pseudo code of the proposed PSO classifier is as follows:

```
For each particle
   Initialize particle
END
Do
   For each particle
      Calculate fitness value
      If the fitness value is better than the best fitness value
(pBest) in history
set current value as the new pBest
   End
   Choose the particle with the best fitness value of all the
particles as the gBest
   For each particle
      Calculate particle velocity according equation (a)
      Update particle position according equation (b)
   End
While maximum iterations or minimum error criteria is not
attained
```

Particles' velocities on each dimension are fixed to a maximum velocity Vmax. If the sum of accelerations would

cause the velocity on that dimension to exceed Vmax, which is a parameter specified by the user. Then the velocity on that dimension is limited to Vmax.

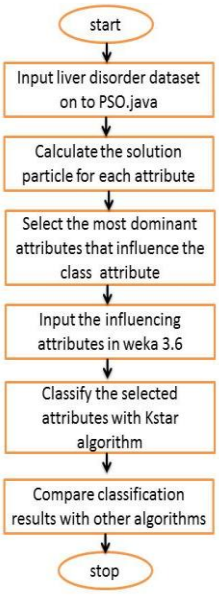

Fig 1. Flowchart of the proposed work

This paper emphasize on the implementation of PSO over any data set, which would be helpful for selecting the most influencing attribute for the getting the best classification results. Hence, in this paper, PSO algorithm is designed for attribute selection and then the selected attributes are classified using KStar algorithm over WEKA tool.

IV. EXPERIMENTATION

A dataset contains the details of patients' with liver disorder is taken from the UCI machine learning repository [17]. The dataset holds 345records with seven attributes including the selector (classification) attribute. The first 5 variables are all blood tests which are thoughtto be sensitive to liver disorders that might arise fromexcessive alcohol consumption. Each record in the file constitutes the record of a single male individual. The selector attributes determines the presence of the liver disorder. The initial particles are generated for training data to retrieve the pbest and gbest, the pbest are then compared with the lbest to set the best solution for attribute selection. The proposed PSO algorithm is implemented using a java application. The dataset is executed with the PSO java program for attribute selection. The description of the data set is explained in Table 1.

Table 1. Dataset Description

S.No	Attribute	Description
1	Mcv	mean corpuscular volume
2	Alkphos	alkaline phosphatase
3	Sgpt	alamine aminotransferase
4	Sgot	aspartate aminotransferase
5	Gammagt	gamma-glutamyltranspeptidase
6	Drinks	number of half-pint equivalents of alcoholic beverages drunk per day
7	Selector	field used to classify into sets

V. RESULTS AND DISCUSSION

PSOClass.java program identified solution particles for all attributes in the dataset. Out of six attributes four attributes have been recognized as the most influencing attributes in yielding cent percent classification result over liver dataset such as alkphos, sgpt, sgot, gammagt. The least influencing attributes are mcv and drinks which have produced only 72% classification results. The results of PSOClass.java and the solution particles of the data set is depicted in Fig.2 and Table.2 respectively. The selected attributes are then inputted onto the WEKA tool [18] for performing classification. The algorithm used for classifying the attributes is KStar algorithm.

Table 2. PSO- Solution Particle Summary of Individual Attributes

Attribute Name	PSO Value
Gammagt	4.605919250169462
Alkphos	4.497428985573608
Sgpt	3.91061372180219
Sgot	3.076856854666527
drinks	1.3646868217604802

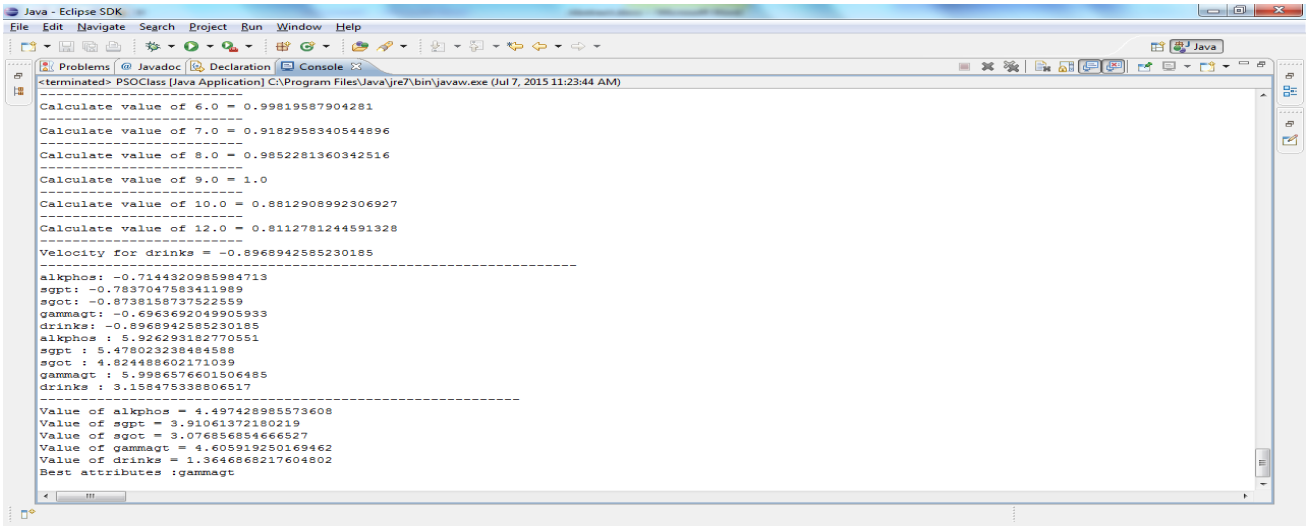

Fig.2 PSOClass.java Solution Particles

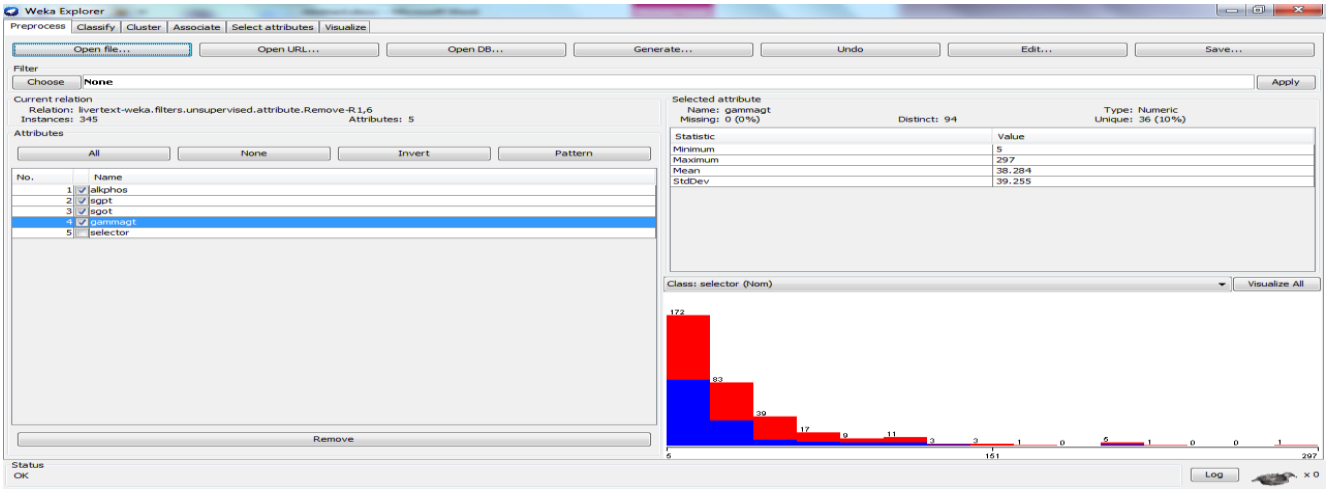

Fig 3. Best PSO_Attributes

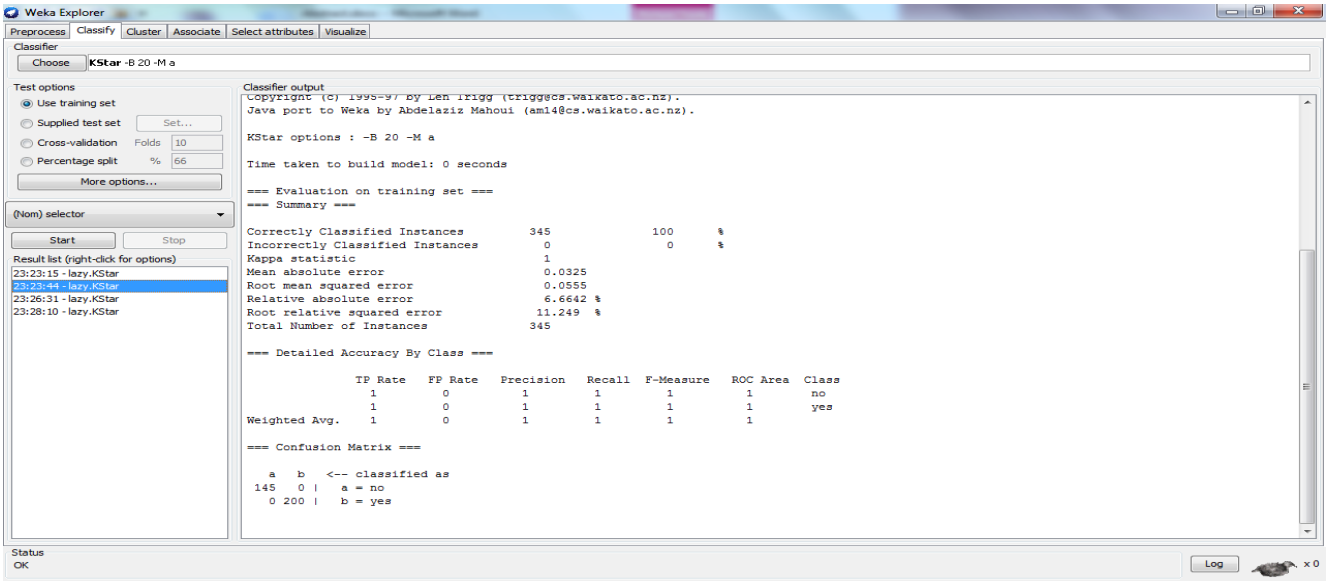

Fig 4. Classification result of Best PSO_Attributes using KStar

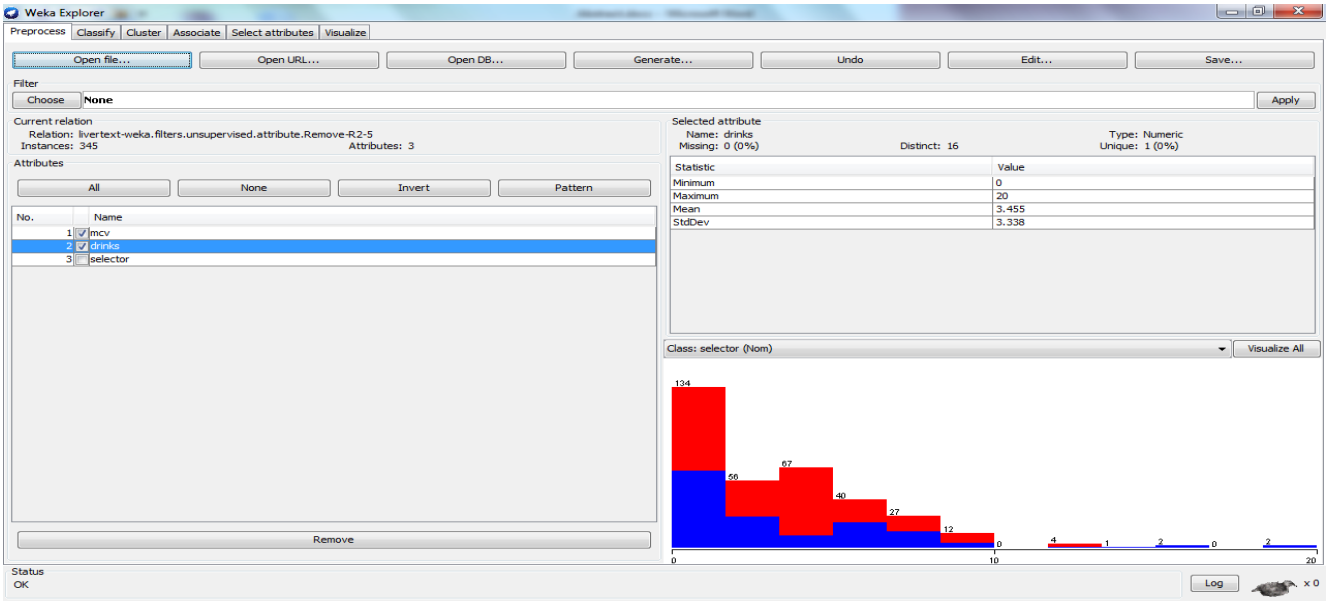

Fig 5 Least PSO_Attributes

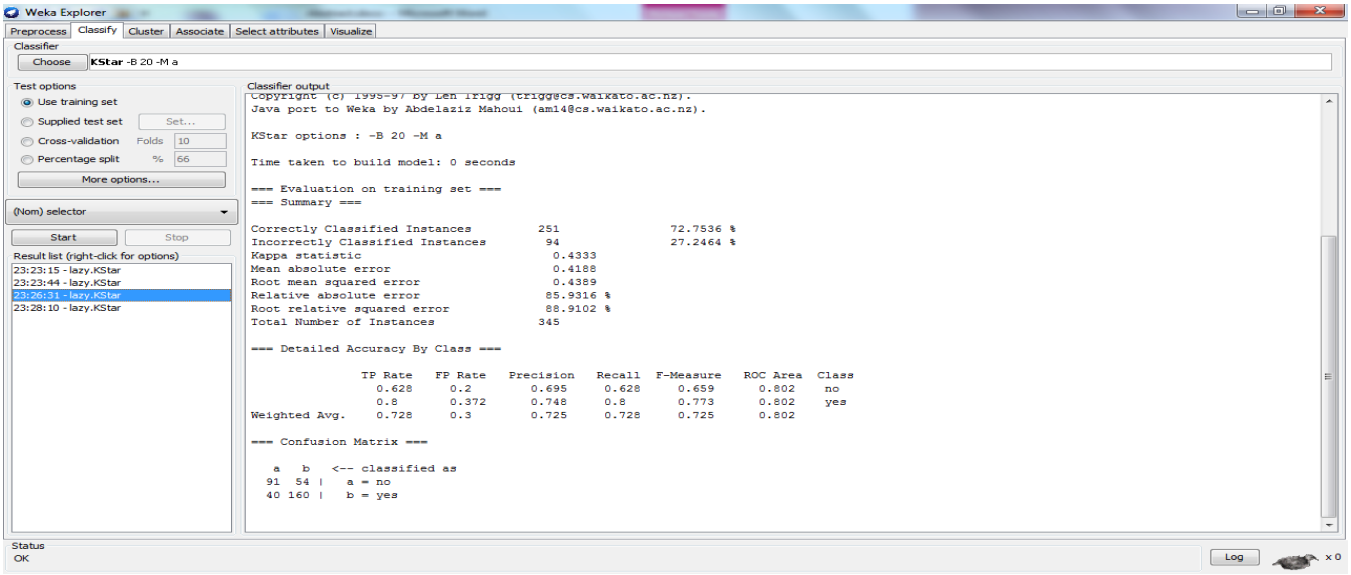

Fig 6. Classification result of Least PSO_Attributes using KStar

Moreover, the performance of PSO Selector is tested with other classification algorithms to experience the accuracy resulting by them. Unlike, The Kstar algorithm can be defined as a method of classification analysis which mainly aims at the partition of n observation into k clusters in which each observation belongs to the cluster with the nearest mean. Kstar algorithm is described as an instance based learner which uses entropy as a distance measure. The benefits are that it provides a consistent approach to handling of real valued attributes, symbolic attributes and missing values. [12] KStar, most of the algorithms could not provide highest accuracy. The classification accuracy with respect to the PSO Selected Attributes is evaluated and presented in Table 3. The pictorial representation of the comparative analysis of PSO-KSTAR with existing algorithms is depicted in Fig. 7.

Table 3. Comparative Analysis of Pso-Kstar with Existing Classifiers

S.No	Algorithm	Classification Accuracy
1	PSO-KSTAR	100%
2	PSO-NAÏVE BAYES	89%
3	PSO-RBF NETWORK	93%
4	PSO-JRIP	85%
5	PSO-J48	96%

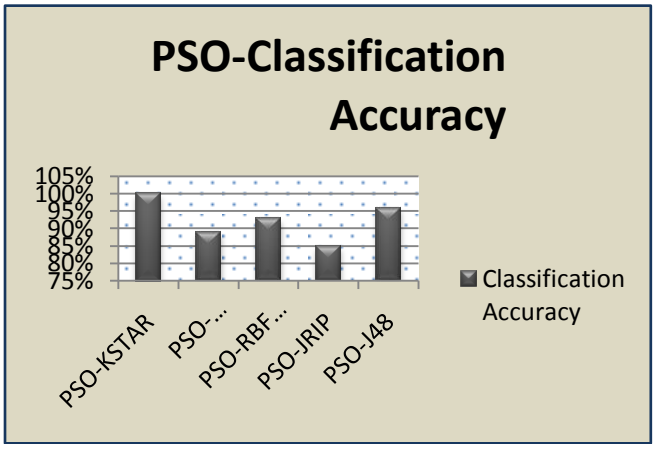

Fig 7. PSO-Kstar Vs Existing Classifiers

VI. **CONCLUSION**

Liver is one of the important internal organs of the human body, and is responsible for more than one hundred internal functions. An issue in this organ may easily lead to liver disorder problems which affects the functions of the liver. So early diagnose of liver disorder may increase the surveillance the patients. PSO-Kstar algorithm is best suitable algorithm for the classification of liver disorders as it improved the performance in prediction accuracy as it is discussed earlier. PSO-Kstar algorithm is considered as one of good data mining algorithm with respect to understandability, transformability and accuracy gives 100% . In future, this algorithm can be applied other diseases.

REFERENCES

[1] K. Senthamaraikannan, N.SenthilvelMurugan, V.Vallinayagam and T. Viveka, *"Analysis of Liver Cancer DNA Sequence Data using Data Mining"* International Journal of Computer Applications (0975 – 8887) Volume 61– No.3, January 2013.

[2] Cheng-Mei Chen, Chien-Yeh Hsu, Cheng-Mei Chen and Chien-Yeh Hsu*," Prediction of Survival in Patients with Liver Cancer using Artificial Neural Networks and Classification and Regression Trees",*Seventh International Conference on Natural Computation 2011.

[3] P.Rajeswari, G.SophiaReena, *"Analysis of Liver Disorder Using Data mining Algorithm",* Vol. 10 Issue 14 (Ver. 1.0), Global Journal of Computer Science and Technology, November 2010.

[4] A.S.Aneeshkumar, C.JothiVenkateswaran, *"Estimating the Surveillance of Liver Disorder using Classification Algorithms"*, International Journal of Computer Applications (0975 – 8887)Volume 57– No.6, November 2012.

[5] HyontaiSug, *"Improving the Prediction Accuracy of Liver Disorder Disease with Oversampling",*Applied Mathematics in Electrical and Computer Engineering,ISBN: 978-1-61804-064-0.

[6] Ziqiang Wang, Xia Sun, and Dexian Zhang, *"A PSO-Based Classification Rule Mining Algorithm",*ICIC 2007, LNAI 4682, pp. 377–384, 2007.

[7] Gopala Krishna Murthy Nookala, Bharath Kumar Pottumuthu, NagarajuOrsu, Suresh B. Mudunuri, *" Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification" ,* (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No.5, 2013.

[8] A.Priyanga and Dr,S.Prakasam *" The Role of Data Mining-Based Cancer Prediction system (DMBCPS) in Cancer Awareness"*, International Journal of Computer Science and Engineering Communications- IJCSEC. Vol.1 Issue.1, December 2013.

[9] J.S.Saleema, N.Bhagawathi, S.Monica, P.DeepaShenoy, K.R.Venugopal and L.M.Patnaik*," Cancer Prognosis Prediction using Balanced Stratified Sampling"*, International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.3, No. 1, February 2014.

[10] ShwetaKharya, *"Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease",*International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April [11].GiduduAnthonya and Heinz Rutherb, "*Comparison of Feature Selection Techniques for SVM Classification*", International Society for Photogrammetry and Remote Sensing, Proceedings, XXXVI/7, 2011.

[11] GiduduAnthonya and Heinz Rutherb, "*Comparison of Feature Selection Techniques for SVM Classification*", International Society for Photogrammetry and Remote Sensing, Proceedings, XXXVI/7, 2011.

[12] Mark A. Hall, "*Correlation-based Feature Selection for Machine Learning*", University of Waikato, 1999

[13] K.Revathi, T.KalaiSelvi, "*Survey: Effective Feature Subset Selection Methods and Algorithms for High Dimensional Data*", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 12, December 2013

[14] AmiraSayed A. Aziz, Ahmad TaherAzar, Mostafa A. Salama, "*Genetic Algorithm with Different Feature Selection Techniques for Anomaly Detectors Generation*", Proceedings of the 2013 Federated Conference on Computer Science and Information Systems pp. 769–774, 2013

[15] BangsukJantawan, Cheng-Fa Tsai, "*A Comparison of Filter and Wrapper Approaches with Data Mining Techniques for Categorical Variables Selection*", International Journal of Innovative Research in Computer and Communication Engineering Vol. 2, Issue 6, June 2014

[16] BinitaKumari, TriptiSwarnkar, "*Filter versus Wrapper Feature Subset Selection in Large Dimensionality Micro array: A Review*", International Journal of Computer Science and Information Technologies, Vol. 2 (3) , 2011, 1048-1053 2012

[17] https://archive.ics.uci.edu/ml/datasets.html

[18] www.cs.waikato.ac.nz/ml/weka/