

SS-SVM (3SVM): A New Classification Method for Hepatitis Disease Diagnosis

Mohammed H. Afif, Abdel-Rahman Hedar, Taysir H. Abdel Hamid, Yousef B. Mahdy
Faculty of Computers & Information
Assiut Univ.
Assiut 71526, EGYPT

Abstract—In this paper, a new classification approach combining support vector machine with scatter search approach for hepatitis disease diagnosis is presented, called 3SVM. The scatter search approach is used to find near optimal values of SVM parameters and its kernel parameters. The hepatitis dataset is obtained from UCI. Experimental results and comparisons prove that the 3SVM gives better outcomes and has a competitive performance relative to other published methods found in literature, where the average accuracy rate obtained is 98.75%.

Keywords—Support Vector Machine; Scatter Search; Classification; Parameter tuning

I. INTRODUCTION

The classification problem may be encountered in different domains, such as "disease diagnosis". Disease diagnosis usually depends on many symptoms and results of medical exams that demonstrate the presence or absence of the disease. Thus, disease diagnosis can be described as a classification problem.

Recently, many researchers try to propose new classification methods to improve or enhance the outcomes of existing methods. Several machine learning algorithms and data mining tools are employed; most studies are interested in proposing new methods that may be help in diseases diagnosis. The term hepatitis means an inflammation of the liver without determining a specific reason [28], [6]. There are more than two viruses that cause hepatitis, the serious of them are Hepatitis B virus (HBV) and Hepatitis C virus (HCV), where about 600000 and more than 350000 people died every year from HBC and HCV, respectively according to WHO (World Health Organization) statistic, Also, Countries with high rate from (HCV) are Egypt (22%), Pakistan (4.8%) and China (3.2%) [1]. This study concentrates on hepatitis disease due to its wide spread diseases around the world, as well as proposing a new method that may help the diagnosis of this serious disease.

The suggested method 3SVM combined support vector machine with scatter search (SS) approach. The SVM algorithm is used due to the following advantages: SVMs one of the most powerful classifiers and is applied to many different domains like pattern recognition [5], and bioinformatics [27], in the case separable datasets SVMs can find the optimal separation hyperplane, SVMs have ability to deal with very high dimensional data " means handle the curse of dimensionality well" [33], from computation perspective SVMs provide a fast training. Furthermore, SS methodology is

employed due to its promising performance when applied with machine learning algorithms.

Hepatitis datasets used are obtained from UCI repository. The main difference with other methods published in literature is the usage of SS approach to find near optimal values of SVM parameters and its kernel parameters. All features of the datasets are used without applying any reduction techniques, using SVMs classifier, in addition, two types of experiments are conducted using 10-fold cross validation method and holdout method for datasets partitioning with three different rates (50-50%,70-30%,80-20%), for training and testing, respectively. The obtained results are very promising where the accuracy is 98.75% in case of 10-fold method, while 92.5%, 95.83% and 100% for the three partition methods, respectively.

The paper is organized as follows. Next section gives an overview about the methods that are found in literature. Section 3 gives a brief description about datasets. Section 4 describes the 3SVM steps in details. Section 5 reports numerical experiments and results. Finally, the conclusions and future work make up Section 6

II. RELATED WORK

This section summarizes some works that found in literature. Plot and Günes in [28] present a new method called FS-AIRS with fuzzy resource allocation for hepatitis diagnosis. The method relies on a hybrid method that uses Feature Selection (FS) and Artificial Immune Recognition System (AIRS) with fuzzy resource allocation mechanism. The obtained results are very promising when compared with more than 20 approaches proposed in literature, where the average accuracy rate is 92.59%. Authors in [14] suggest a new method for classifying medical data, where a hybrid model is proposed by combining a case-based data clustering method and a fuzzy decision tree. The model is tested by using breast cancer wisconsin (diagnosis) and liver disorders datasets from UCI, where the produced accuracy rate is 98.4% and 81.6%, respectively. Researchers conclude that the proposed method could help doctors to extract effective conclusions in medical diagnosis. In [16], a new classification approach called FCS-ANTMINER is presented, where ant colony optimization is used to extract a set of fuzzy rules for diagnosis of diabetes disease; the accuracy rate is 84.24%. Researchers in [7] present a new hybrid method called LFDA-SVM for hepatitis disease diagnosis; Local Fisher Discriminant Analysis (LFDA) and SVM are combined. The method employed LFDA for performing feature reduction to improve the performance of

standard SVM algorithm. Datasets from UCI is used in testing, and the obtained accuracy rate is (96.77%) which is the best results when compared with other published approaches in literature. Also, a new intelligent method for hepatitis disease diagnosis called PCA-LSSVM, is suggested by [6]. The proposed method based on Principle Component Analysis (PCA) and Least Square SVM (LSSVM).The PCA is employed for feature extraction and reduction while LSSVM for classification, using Hepatitis datasets from UCI repository. The accuracy rate that produced is(95%). Furthermore, authors in [32] present a hybrid method based on SVM combined with Simulated Annealing (SA) for hepatitis disease diagnosis, also the method uses the same datasets which is used by previous studies in [7],[6]. The obtained accuracy is (96.25%), which the best accuracy rate when compared with other methods. In recent study presented by [4], the authors summarized the most works in the area of hepatitis disease diagnosis, and proposed a new method by employing Probabilistic Neural Network structure called PNN (10xFC), the results that obtained is 91.25%.

Classification results of the most previous methods may need to be enhanced or improve, especially when applied to critical applications, such as disease diagnosis. The diagnosis of some disease like hepatitis is very difficult task for a doctor, where doctor usually determines decision by comparing the current test results of patient with other one who has the same condition. All these reasons motivated for suggesting new methods to improve the outcomes of existing approaches, as well as to help a doctors and specialists to diagnose hepatitis diseases.

III. DESCRIPTION ABOUT DATASET

This study conducts experiments on hepatitis dataset, which is obtained from UCI machine learning repository. The dataset contains 155 instances distributed between two classes die with 32 instances and live with 123 instances. There are 19 features or attributes, 13 attributes are binary while 6 attributes with 6-8 discrete values. The goal of the dataset is to forecast the presence or absence of hepatitis virus. Table I lists information about the features.

TABLE I. INFORMATION ABOUT THE FEATURES OF THE HEPATITIS DATASET

Number	Name of features	The values of features
1	Age	10,20,30,40,50,60,70,80
2	Sex	Male, Female
3	Steroid	Yes, No
4	Antivirals	Yes, No
5	Fatigue	Yes, No
6	Malaise	Yes, No
7	Anorexia	Yes, No
8	Liver big	Yes, No
9	Liver firm	Yes, No
10	Spleen palpable	Yes, No
11	Spiders	Yes, No
12	Ascites	Yes, No
13	Varices	Yes, No
14	Bilirubin	0.39,0.80,1.20,2.00,3.00,4.00
15	Alk phosphate	33,80,120,160,200,250
16	SGOT	13,100,200,300,400,500
17	ALBUMIN	2.1,3.0,3.8,4.5,5.0,6.0
18	PROTIME	10,20,30,40,50,60,70,80,90
19	HISTOLOGY	Yes, No

IV. THE PROPOSED METHODOLOGY

a) In this section, SVM and its parameters are defined. In addition, the steps of 3SVM are explained in details, as illustrated in figure

b) Support Vector Machine and Solution Definition

Support Vector Machines (SVMs) one of the promising machine learning algorithms, which depends on statistical learning theory developed by Vapnik [10, 35, 11, 19, 2]. The main problems that are encountered in SVMs are how to find near optimal values for its parameters and select a SVMs kernel as well as tuning its parameter. The parameters that should be optimized are the complexity parameter C , epsilon ϵ and tolerance t and the kernel function parameters, such as γ for Gaussian kernel. The parameter C determines the trade-off between the fitting error minimization and model complexity [37, 29, 9, 24], where a bad choice of C leads to an imbalance between model complexity minimization and empirical risk minimization. The last two parameters ϵ , where its value indicates the error expectation in the classification process of the sample data, and it impacts the number of support vectors generated by the classifier [24], while t , is the tolerance parameter. In 3SVM the solution for finding the near optimal values of SVMs parameters and its kernel is represented as vector with dimensions equal to the number of trial solutions as in equation 1.

$$X = [P_1, P_2, P_3, P_4] \quad (1)$$

Where P_1 is kernel parameter in range [0.0001, 33], while others are SVM parameters P_2 C is Complexity and its range [0.1, 35000], P_3 ϵ is epsilon [0.00001, 0.0001] and P_4 t tolerance [0, 0.5]. These chosen values are based on the common settings in the literature [12, 36, 8]. As known, the classification process is divided into two phases: model building and model testing. In first phase, a learning algorithm runs over datasets to develop a model that could be employed in estimating an output. The aim of the model is to describe the relationship between the class and the predictor [15, 20, 13, 30]. The quality of the produced model is assessed in the model testing phase. Usually, accuracy measure is used for assessing the performance of the most classification methods, where it is calculated as in equation 2.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

where, TP (True Positive) is the positive cases that are classified correctly as positive, TN (True Negative) is the negative cases that are classified correctly as Negative, while, FP (False Positive) are cases with negative class classified as positive, and FN (False Negative) is the cases with positive class classified as negative [19, 21, 31]. Thus, the accuracy rate is used in this method to measure the quality of generated solutions, which is called the fitness function (fit). Furthermore, there are other performance measures employed, such as sensitivity. Sensitivity is the proportion of the cases with positive class that are classified as positive (true positive rate, expressed as a percentage); specificity is the proportion of cases with negative rate class, classified as negative (true negative rate, expressed as a percentage). Sensitivity and

specificity reflect how well the classifier discriminates between case with positive and with negative class [19, 21]. They are calculated as in 3 and 4 equations as below:

$$\text{Sensitivity} = \frac{TP}{(TP+FN)} \quad (3)$$

$$\text{Specificity} = \frac{TN}{(TN+FP)} \quad (4)$$

c) Preprocessing Phase

To use SVM classifier all features of the datasets must be set in real number format. Therefore, the nominal features are converted into numerical data. After that, data normalization using equation 5 is performed. In order to prevent feature values in greater numeric ranges from dominating and to avoid numerical difficulties during the calculation. In addition, two methods are used in splitting dataset for training and testing phase. In first method is k-fold Cross Validation (CV), which is a popular strategy to estimate the performance of the classification methods, as well as to avoid trap in over-fitting problem, where the training sample is independent from the testing sample [3, 19, 2]. In k-fold CV the k value is usually set to 10. Therefore, the datasets are split into 10 parts. Nine data parts are applied in the training process, while the remaining one is utilized in the testing process. The program is run 10 times to enable each slice of data to take a turn as testing data. The accuracy rate for training process and testing process is calculated by summing the individual accuracy rates and error rates for each time of run, and then divided by 10. The second method is holdout. The datasets are split into two parts: one for training and the second for testing with various rates 50% - 50%, 70% - 30%, 80% - 20%, respectively. The major aim from using two methods for dataset split is evaluate the applicability of the method from more than one perspective.

$$X_{\text{Normalization}} = \frac{X - X(\min)}{X(\max) - X(\min)} \quad (5)$$

d) Applying Scatter Search Methodology

Scatter Search (SS) is one of meta-heuristics approaches classified as population-based algorithm, which is first suggested by F. Glover in the 1970's [18], due to his results in 1960's [17]. SS has more flexible framework than the other Evolutionary algorithms and uses a memory-type diversification procedure for more efficient globally search [22]. In addition, Glover in 1998 [26, 22] published the SS template, which presents an algorithmic description of the SS method. In addition, the SS is a promising meta-heuristic technique and has been applied to many different applications successfully. Some of these applications may be found in [22]. Furthermore, there are some studies that applied SS to machine learning algorithms as in [34], Authors suggest a hybrid procedure combining neural networks, and scatter searches to optimize the continuous parameter design of back-propagation neural network. Another method is suggested by [25] which constructed three scatter search-based algorithms to solve the feature-selection problem. In the area of parameter setting, a few works are done using SS. Lin *et al.* in [23] suggest an approach to determine the parameters and feature selection for C4.5 algorithm by employing SS meta-heuristics strategy. In

[8], researchers propose a method to enhance the classification accuracy by using SS approach to determine the parameters of three machine learning algorithms and performing feature selection for these algorithms. The SS depends on five steps which are:

Diversification Generation Method, An improvement Method, Reference Set Method, Subset Generation Method and Solution Combination.

1)- Diversification Generation Method:

After the preprocessing phase, the first method of scatter search is invoked in which a set of random solutions (value for parameters) are generated, based on the upper and, lower bound of every parameters defined in first section, and according to equation 6, the number of generated solutions is 30.

$$Sol_x = LWB[i] + (UPB[i] - LWB[i]) \times Rnd \quad (6)$$

where the $LWB[i]$: is the Lower Bound of the parameter number i , UPB : is the Upper Bound of the parameter number i and Rnd : is a random value in (0,1). After that, the model is training and testing using all solutions that are generated. After that, the initial Reference Set (*RefSet*) is develop by selecting the b solutions that produce the best accuracy rate $b=5$. After that, the subset generation, solution combination and *Refset* update steps, as described below, are repeated until one of the termination conditions is satisfied. This paper defines three termination conditions and if any condition of them is satisfied the process will be stopped. The conditions are: - **First Condition**: When the accuracy rate gets up to 100% for at least one solution.

-**Second Condition**: When $I_{max} \geq 75$, where I_{max} is the maximum iteration.

-**Third Condition**: When $OldRefset = NewRefSet$ this means that no update is achieved.

2)- Subset Generation Method:

This method depends on or operates on the *RefSet* to generate all pairs of solutions, where the maximum number of subsets is $(b^2-b)/2$. Means that 10 subsets are generated each subset is pair of solutions.

3)- Solution Combination Method:

In this method, a number of new solutions are generated from each subset of parents P_1 and P_2 as follows:

$$X_1 = P_1 + (P_2 - P_1) \times r_1 \quad (7)$$

$$X_2 = P_1 + (P_2 + P_1) \times r_2 \quad (8)$$

$$X_3 = P_1 + P_2 \times r_3 \quad (9)$$

Where r_1 , r_2 and r_3 are random numbers in (0,1). From this method, there are 30 new solutions are generated these solution will be used for training and testing the model. After that, solutions are put in pool together with solutions in the *Refset* in order from the best one to worst.

4)- Solution Combination Method:

In this method, the *Refset* is updated to has the best $b_1=4$ solutions from the pool and the $b_2=1$ diverse solutions, where $b_1+ b_2 = b$. Diverse solution is selected, which depends on calculating the Euclidean distance for each solution in the *Refset* and solutions in the pool. The b_2 solution with the maximum distance is selected as diverse one.

V. NUMERICAL EXPERIMENTS

The 3SVM approach is implemented on PC with Core2Duo 2.93 Ghz CPU, 2GB of RAM, and windows XP Professional OS. Visual Studio 2008-Visual C# and Accord.net framework are used in development.

A. Results and Discussion

The 3SVM approach performs two types of experiments: first one: uses k-fold cross validation method for splitting the dataset, and the second holdout method is used. Tables II and III list results of experiments, which are produce by using two different range for parameter C as in tables. Table II contains first row accuracy rate for testing (**ACC.TS**), and the remainder rows contained standard deviation for accuracy of testing (**STDEV.TS**), accuracy rate for training (**ACC. TR**), rate for training (**ERR.TR**), standard deviation for error rate of training phase (**STDEV.Err.TR**), sensitivity and specificity. While Table III contains in the first row the number of generation when the best is obtained (**No.Gen.Best Sol.Obt.**) and the rest of the rows contain the number of hitting the best solution (**No.Hit.Best Sol.**) and fitness function evaluation times (**Fitness Fun.Eval**). These factors reflect some performance aspects of the 3SVM method. All obtained results are very promising for various methods of experiments. The obtained accuracies are **98.75%**, **93.75%**, **91.66%** and **87.5%** for first range and **98.75%**, **100%**, **95.83%** and **92.5%** for the second range. The best results appear when using second range [0.1, 35000] for different methods of splitting dataset. In addition, figure 2 displays the accuracy rate for training data and testing data of various methods for splitting dataset. The differences between the accuracy rate for training and testing are reasonable for all splitting methods. This proves that the 3SVM method does not suffer from over-fitting and under-fitting problems, according to the fact that there is no large difference between the training and testing accuracy [23, 8]. Furthermore, the classification outcomes of 3SVM approach are compared with results of other published approaches. Table IV lists comparisons of **30** methods proposed in literature as listed in [6], [7], [23] and [4]. From comparisons, the 3SVM gives the better results than other methods proposed in literature, where the 3SVM enhances the performance of classification and the accuracy rate increases with **2.5%**, **1.98%** and **7.5%** from the recently published methods [32], [7] and [4]. In addition, there are some major differences with other approaches in literature like some methods perform feature reduction, as well as using different training algorithms like neural network, using different implementation environments and different tools for SVM implementation. Finally, one may conclude that obtained results by 3SVM method is encouraged and gives the best performance when compared with methods that are published recently, [4], [6], [7] and [32], as summarized in Table IV. In addition, the

experimental results prove the efficiency of scatter search method for tuning SVM parameters. Therefore, the 3SVM method may be successfully employed to help doctors or specialists in diagnosis of hepatitis disease, providing them with some hints or indication that may help in making decision for disease diagnosis.

TABLE II. RESULTS

Measure	Rang of C parameter	The Method Used			
		10-fold	80-20%	70-30%	50-50%
ACC.TS	0.1-25000	98.75	93.75	91.66	87.5
STDEV.TS		0.0395	-	-	-
ACC.TR		99.16	100	100	100
ERR.TR		0.00833	0	0	0
STDEV.TR		0.0134	-	-	-
Sensitivity		100	100	95	96.87
Specificity	30	0	75	50	
ACC. TS	0.1-35000	98.75	100	95.83	92.5
STDEV.TS		0.0395	-	-	-
ACC. TR		99.86	100	100	95
ERR.TR		0	0	0	0.05
STDEV.TR		0.00439	-	-	-
Sensitivity		98.57	100	95.45	100
Specificity	80	100	100	0	

TABLE III. RESULTS

Measure	Rang of C Parameter	The Method Used			
		10-fold	80-20%	70-30%	50-50%
No.Gen.Best Sol.Obt	0.1-25000	89	75	75	75
No.Hit.Best Sol.		286	0	0	0
Fitness Fun.Eval.		3410	2650	2650	2650
No.Gen.Best Sol.Obt	0.1-35000	82	2	75	75
No.Hit.Best Sol.		118	1	0	0
Fitness Fun.Eval.		3170	100	2650	2650

TABLE IV. RESULTS OF 3SVM

COMPARED WITH APPROACHED [6],[7],[32],[4],[7]

Author	Method	Accuracy
Adamczak	MLP+BP(Tooldiag)	77.4
Adamczak	RBF(Tooldiag)	79.0
Adamczak	FSM with rotations	89.7
Adamczak	FSM without rotations	88.5
Bascil and Temurtas	MLNN(MLP) + LM	91.87
Bascil and Oztekin [4]	PNN(10×FC)	91.25
Chen,Liu, <i>et al</i> [7]	L DFA-SVM	96.77
Calisir and Dogantekin [6]	PCA-LLSSVM	95
Dogantekin, Avci	LDA-ANFIS	94.16
Grudzinski	Weighted9NN	92.9
Grudzinski	18NN,stand.Manhattan	90.2
Grudzinski	15NN,stand.Euclidean	89.0
Jankowski	IncNet	86.0
Ozyildirim, Yildirim <i>et al</i>	MLP	74.37
Ozyildirim, Yildirim <i>et al</i>	RBF	83.75
Ozyildirim, Yildirim <i>et al</i>	GRNN	80.0
Polat and Gunes	FS-AIRS with fuss res	92.59
Polat and Gunes	PCA-AIRS	94.12
Stern and Dobnikar	LDA	86.4
Stern and Dobnikar	NaiveBayes and semi-NB	86.3
Stern and Dobnikar	QDA	85.8
Stern and Dobnikar	1-NN	85.3
Stern and Dobnikar	ASR	85
Stern and Dobnikar	FDA	84.5
Stern and Dobnikar	LVQ	83.2
Stern and Dobnikar	CARC (DT)	82.7
Stern and Dobnikar	ASI	82.0
Stern and Dobnikar	LFC	81.9
Stern and Dobnikar	MLP with BP	82.1

Sartakhti, Zangoeei <i>et al</i> [32]	SVM+SA	96.25
Our Method	3SVM	98.75

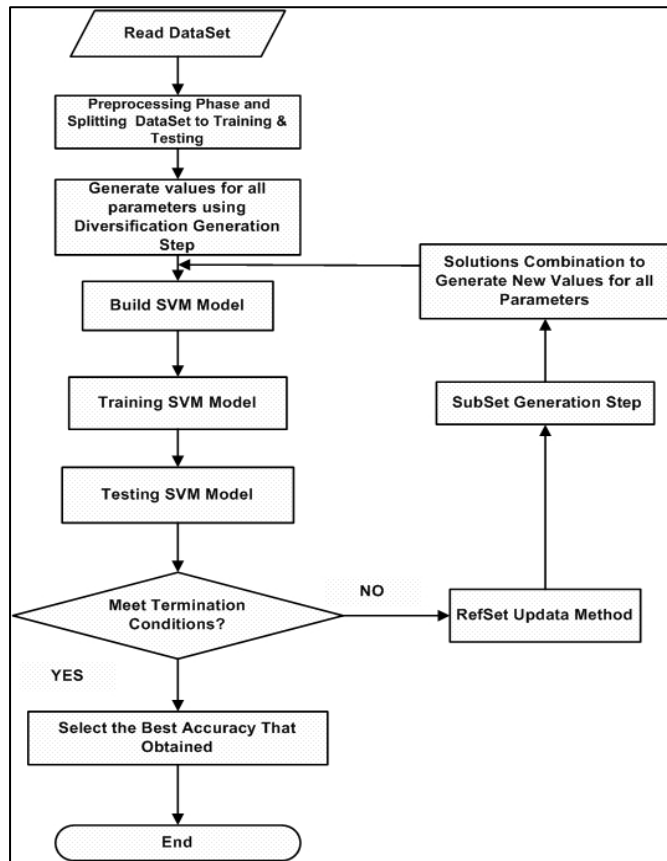


Fig. 1. Methodology Steps Flowchart

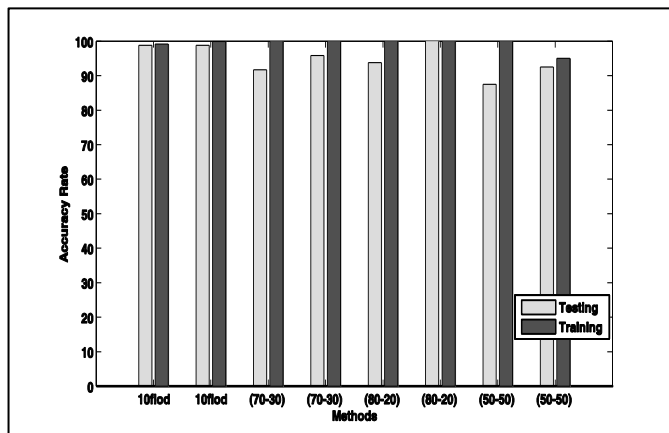


Fig. 2. The accuracy for Training and Testing Process

VI. CONCLUSIONS AND FUTURE WORK

This paper proposed a new method 3SVM, for hepatitis virus diagnosis, which combined SVM with scatter search. Experiments proved that 3SVM has very promising performance in classifying the living liver from the dead one, with accuracy rate **98.75%**. Also, experiments demonstrated that the SS was a practical approach for tuning parameters of

SVM and its kernel parameters. A comparison of the obtained results with other published approaches found in literature demonstrated that the 3SVM given better results than others. However, the 3SVM method may be successfully used to help diagnosis of hepatitis disease. In future, the performance of the proposed method may be enhanced by performing feature reduction. In addition, more features will be added to existing datasets to enhance 3SVM to be able to forecast the treatment procedures according to the level of disease. Moreover, 3SVM will be applied to multiclass problems.

REFERENCES

- [1] World health organization (who). <http://http://www.who.int>, visted in Jul, 15, 2012.
- [2] S. Abe. Support Vector Machines for Pattern Classification. Springer, 2010.
- [3] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4(1):40–79, 2009..
- [4] M.S. Bascil and H. Oztekin. A study on hepatitis disease diagnosis using probabilistic neural network. *MEDICAL SYSTEMS*, 36(3):1603–1606, 2012.
- [5] C.J. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.
- [6] D. Calisir and E. Dogantekin. A new intelligent hepatitis diagnosis system: Pca-lssvm. *Expert Systems with Applications*, 38(8):10705–10708, 2011.
- [7] H-L. Chen, D-Y. Liu, B. Yang, J. Liu, and G. Wang. A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis. *Expert Systems with Applications*, 38(9):11796–11803, 2011.
- [8] S.C. Chen, S.W. Linb, and S.Y. Chou. Enhancing the classification accuracy by scatter-based ensemble approach. *Applied Soft Computing*, p: 1–8, 2010.
- [9] V. Cherkassky and Y. Ma. Pratical selection of svm parameters and noise estimation for svm regression. *Neural Networks*, 17:113–126, 2004.
- [10] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.
- [11] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.
- [12] D. DeCoste and K. Wagstaff. Alpha seeding for support vector machines. *Proceedings of International Conference on Knowledge Discovery and Data Mining*, 2000.
- [13] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*, 2nd ed. John Wiley and Sons Ltd, 2001.
- [14] C-Y. Fan, P-C. Chang, J-J. Lin, and J.C. Hsieh. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. *Applied Soft Computing*, 11(1):632–644, 2011.
- [15] A.H. Fielding. *Cluster and Classification Techniques for the Biosciences*. Cambridge University Press, 2007.
- [16] M.F. Ganji and M.S. Abadeh. A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis. *Expert Systems with Applications*, 38(12):14650–14659, 2011.
- [17] F. Glover. Surrogate constraints. *Operations Research*, 16(4):741 – 749, 1968.
- [18] F. Glover. Heuristics for integer programming using surrogate constraints. *Decision Sciences*, 8(1):156 – 166, 1977.
- [19] L. Hamel. *Knowledge Discovery with Support Vector Machines*. John Wiley and Sons Ltd, 2009.
- [20] F. V. D. Heijden, R.P.W. Duin, D. de Ridder, and D.M.J. Tax. *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*. John Wiley and Sons Ltd, 2004.

- [21] C-L. Huang and C-J. Wang. A ga-based feature selection and parameters optimization for support vector machines. *Expert Systems with Applications*, 31:231–240, 2006.
- [22] M. Laguna and R. Marti. *Scatter search methodology and implementations in C*. Kluwer Academic, 2003.
- [23] S-W. Lin and S-C. Chen. Parameter determination and feature selection for c4.5 algorithm using scatter search approach. *Soft Comput*, 16(1):63–75, 2012.
- [24] S. Liu and N. Jiang. Svm parameters optimization algorithm and its application. In *Proceedings of 2008 IEEE international conference mechatronics and automation*, pages 509 – 513. 2008.
- [25] F.G. Lopez, M.G. Torres, and B.M. Batista. Solving feature subset selection problem by a parallel scatter search. *Operational Research*, 169:477–489, 2006.
- [26] R. Marti, M. Laguna, and F.Glover. Principles of scatter search. *Operational Research*, 169:359–372, 2006.
- [27] W.S. Nobel. Support vector machine applications in computational biology. In B. Schoekkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel Methods in Computational Biology*, pages 71–92. MIT Press, Cambridge, MA, 2004.
- [28] K. Polat and S. Gnes. Hepatitis disease diagnosis using a new hybrid system based on feature selection (fs) and artificial immune recognition system with fuzzy resource allocation. *Digital Signal Processing*, 16(6):889–901, 2006.
- [29] Y. Ren and G. Bai. Determination of optimal svm parameters by using GA/PSO. *Journal of computers*, 5(8):1160–1168, 2010.
- [30] L. Rokach. *Pattern Classification Using Ensemble Methods*. World Scientific Ltd, 2010.
- [31] F. Samadzadegan, A. Soleymani, and R.A. Abbaspour. Evaluation of genetic algorithm for tuning svm parameters in multi-class problems. pages 323 – 327. *CINTI2010 - 11th IEEE international Symposium on Computational Intelligence and Informatics*, 2010.
- [32] J.S. Sartakhti, M.H. Zangoeei, and K. Mozafari. Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA). *computer methods and programs in biomedicine*, -:1–10, 2011.
- [33] R. Stoean, C. Stoean, M. Lupsor, H. Stefanescu, and R. Badea. Evolutionary-driven support vector machines for determining the degree of liver fibrosis in chronic hepatitis c. *Artificial Intelligence in Medicine*, 51(1):53–65, 2011.
- [34] C t. Su, M-C. Chen, and H-L. Chan. Applying neural network and scatter search to optimize parameter design with dynamic characteristics. *Operational Research Society*, 56(10):1132–1140, 2005.
- [35] V.N. Vapnik. *Statistical Learning Theory*. New York John Wiley & Sons., 1998.
- [36] P. Williams, S. Li, J. Feng, and S. Wu. A geometrical method to improve performance of the support vector machine. *IEEE transactions On neural networks*, 18(3):942–947, 2007.
- [37] C-H. Wu, G-H. Tzeng, Y-J. Goo, and W-C. Fang. A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. *Expert Systems with Applications*, 32:397–408, 2007.