ISSN XXXX XXXX © 2017 IJESC

IJESC

Research Article Volume 7 Issue No.6

Data Classification using Multiple Support Vector Machine with Ant Colony Optimization

Geetanjali Patel[1], Prof. Avinash Sharma[2]
PG Scholar[1], Associate Professor[2]
Department of CSE
MITS, Bhopal, India

Abstract:
Dynamic trait evaluation and concept appraisal is chief challenging task in the meadow of brook statistics classification. The permanence of information induced a new quality during categorization development, but the categorization procedure is predefined payment for assigning statistics into course group. Stream data comes into numerous feature sub-set format into immeasurable length. The infinite extent not decided the how numerous classes are assigned. Ant Colony Optimization controls the self-motivated feature evaluation development and decreases the possibility of bewilderment in selection of class and increase the classification relative amount of support vector machine. The optimized feature reduces the unclassified region of class during classification. The proposed method for watercourse data classification is MSVM-ACO is implemented in MATLAB R2013a and test the legalization development used some reputed facts set from UCI machine learning prosperity. These data are corpus, forest and finally used glass dataset.

Index terms: multiple support vector machines, ant colony optimization, UCI machine learning prosperity, features

1. INTRODUCTION

Data classification is more challenging than classifying static data because of several unique properties of data streams. First, data streams are assumed to have definite length, which makes it impractical to store and use all the historical data for training. Therefore, traditional support vector learning algorithms are not directly applicable to data streams. Second, data streams observe concept-drift, which occurs when the underlying concept of the data changes over time. In order to attend to concept drift, a classification replica must continuously get a feel for itself to the most current concept. Third, data streams also scrutinize concept evolution, which occurs when an original class appears in the tributary. With the advent of advanced data classification technologies [1], we are able to continuously collect large amounts of data in various application domains, e.g., daily fluctuations of stock market, trace of dynamic processes, credit card transactions, web click stream, networks traffic monitoring, position updates of moving objects in location-based services and text streams from news etc [2]. Due to its potential in industry applications, data stream mining has been studied intensively in the past few years. The general approach is to first learn one or multiple classification models from the past records of the evolving data, and then use a selected model that best matches the current data to predict the new data records. All the existing data classification techniques assume that at each time stamp there are both large amounts of positive and negative training data available for learning. The goal of data classification is to learn a model from past labeled data, and classify future instances using the model. There are many challenges in data stream classification. Data classifiers may either be single model incremental approaches, or ensemble techniques, in which the classification output is a function of the predictions of different classifiers. Ensemble techniques have been more popular than their single model counterparts because of their simpler implementation and higher efficiency.

II. RELATED WORK

Big Data apprehension large-volume, composite, growing fact sets with several, self-sufficient sources. With the speedy expansion of networking, data luggage compartment, and the figures collection aptitude Big Data are now swiftly expanding in all knowledge and engineering domains, including corporal, natal and biomedical sciences. (Xindong; 2014) As the web has emerged as a huge distributed information warehouse, individuals and organizations have been talented to exploit the low charge information and data on the Internet when making commerce decisions. (Kaile; 2006) Numerous huge organizations have various databases distributed in unusual branches, and as a result multi-database mining is an imperative mission for data mining. (Kaile; 2006) Online streaming attribute choice, in which the extent of the quality set is unknown, and not all features are to be had for learning while leaving the amount of observations invariable. (Xindong, 2010) Hybrid algorithm is proposed to crack combinatorial optimization trouble by using Ant Colony and Genetic programming algorithms. (Nada; 2009) Another Bayesian network is learnt at the innermost site using the data transmitted from the slight situate. (Siva; 2001) Rough set conjecture provides a practical mathematical perception to copy useful decisions from authentic life data involving vagueness, hesitation and impreciseness and is therefore applied fruitfully in the meadow of outline recognition, mechanism learning and acquaintance discovery. (Amit; 2014)

III. METHODOLOGY

The algorithm of proposed technique is performed below:
[classified_data] = MSVM_ACO (dataset)
Step 1: The local gradient values are measured at the choose scale-space in the Region of Interest (ROI) around all features in dataset [4].

Step 2: Above phase are performs in iterative form, then all features of data are saved, Now perform ACO technique on data repository for classify data. The orders of phases for search the best features node through ACO technique as below.

2.1 Ant Colony Optimization: The main goal of ACO is to create scientific systems for optimization, and not to create precise technique of nature. The basic scheme of ACO is as below:
2.1.1 Generate ants and states in particular area.
2.1.2 Select next side probabilistically according to the attractive and visibility.
2.1.3 Probability is calculate as follows

$$\Pr = \frac{r(e).\eta(e)}{\sum_{available,edges,e'} \tau(e').\eta(e')}$$

2.1.4 Every ant maintains a register of infeasible transforms for that repetition.
2.1.5 Modify attractive of a side as per to the number of ants that pass through the value of pheromone is update as follows

$$\boxed{\tau(e) := \begin{cases} (1-\rho)\cdot\tau(e), & \textit{if edge is not traversed} \\ (1-\rho)\cdot\tau(e) + \textit{new pheromone}, & \textit{if edge is traversed} \end{cases}}$$

Where argument $0 \le \rho \le 1$ is known as evaporation tempo, Basically Pheromones is equal to max-term storage of an ant colony and following conditions is satisfied.
ρ small → min evaporation → slow evaporation
ρ large → max evaporation → fast evaporation
2.2 Now generate the optimize dataset of feature point.

Step 3: Train the support vector machine network with training feature set.

Step 4: Search the class label of query data through support vector machine.

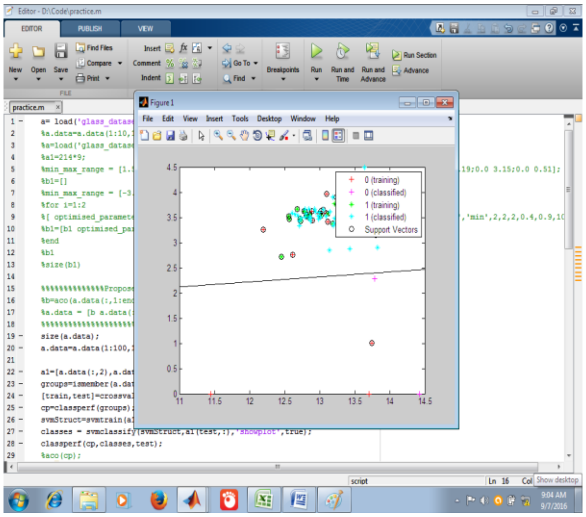
Step 5: Search all relevant data of the same category.

Step 6: Search the adjacent matching data to the query data with the class of data using easy distance metric.

Step 7: Perform support vector machine in iterative label for find actual classified dataset.

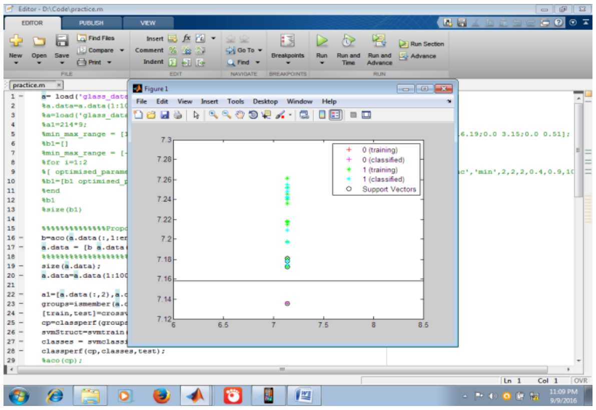
IV. IMPLEMENTATION STEP

The proposed method implements in MATLAB R2013a and tested with very reputed data set from UCI machine learning research center. In the research work, I have measured CR (Correct Rate), ER (Error Rate), PPV (Positive Predictive Value), NPV (Negative Predictive Value), PL (Positive Likelihood) and NL (Negative Likelihood) of classification. To evaluate these performance parameters, It has been used three datasets from UCI machine learning repository namely glass dataset, banana dataset and forest fire dataset. Out of these three dataset, one is small dataset namely glass dataset and other one is large dataset namely as banana dataset. Consider glass dataset input the chunk size as 100 then following window has been displayed.



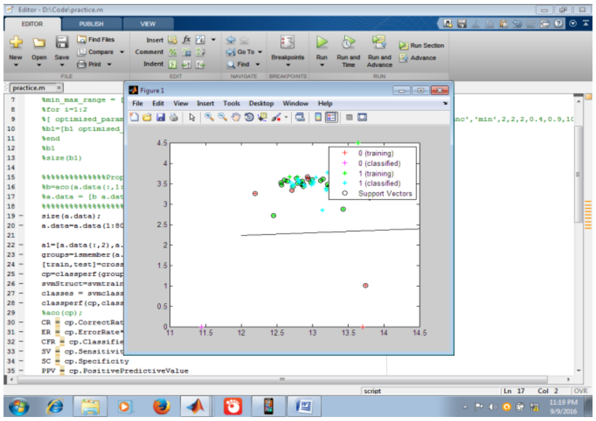
```
a= load('glass_datase
%a.data=a.data(1:10,1
%a=load('glass_datase
%a1=214*9;
%min_max_range = [1.5                    .19;0.0 3.15;0.0 0.51];
%b1=[]
%min_max_range = [-3.
%for i=1:2
%[ optimised_paramete                   ','min',2,2,2,0.4,0.9,10
%b1=[b1 optimised_par
%end
%b1
%size(b1)

%%%%%%%%%%%%%%%Propose
%b=aco(a.data(:,1:end
%a.data = [b a.data(:
%%%%%%%%%%%%%%%%%%%%%%
size(a.data);
a.data=a.data(1:100,1

a1=[a.data(:,2),a.dat
groups=ismember(a.dat
[train,test]=crossval
cp=classperf(groups);
svmStruct=svmtrain(a1
classes = svmclassify(svmStruct,a1(test,:),'showplot',true);
classperf(cp,classes,test);
%aco(cp);
```

Figure.1. Data Classification using Support Vector Machine method for the given chunk size 100 in glass dataset.



```
a= load('glass_data
%a.data=a.data(1:10
%a=load('glass_data
%a1=214*9;
%min_max_range = [1                     .6.19;0.0 3.15;0.0 0.51];
%b1=[]
%min_max_range = [-
%for i=1:2
%[ optimised_parame                     c','min',2,2,2,0.4,0.9,10
%b1=[b1 optimised_p
%end
%b1
%size(b1)

%%%%%%%%%%%%%%%Propo
b=aco(a.data(:,1:en
a.data = [b a.data(
%%%%%%%%%%%%%%%%%%%%
size(a.data);
a.data=a.data(1:100

a1=[a.data(:,2),a.d
groups=ismember(a.d
[train,test]=crossv
cp=classperf(groups
svmStruct=svmtrain(
classes = svmclassi
classperf(cp,classes,test);
%aco(cp);
```

Figure.2. Data Classification using Multiple Support Vector Machine with Ant Colony Optimization method for the given chunk size 100 in glass dataset



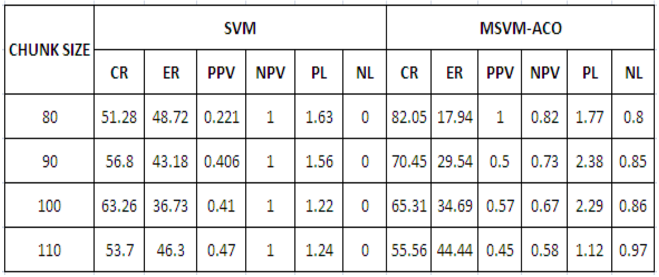
```
%min_max_range = [
%for i=1:2
%[ optimised_param                      nc','min',2,2,2,0.4,0.9,10
%b1=[b1 optimised_
%end
%b1
%size(b1)

%%%%%%%%%%%%%%%Prop
%b=aco(a.data(:,1:
%a.data = [b a.dat
%%%%%%%%%%%%%%%%%%%
size(a.data);
a.data=a.data(1:80

a1=[a.data(:,2),a.
groups=ismember(a.
[train,test]=cross
cp=classperf(group
svmStruct=svmtrain
classes = svmclass
classperf(cp,class
%aco(cp);
CR = cp.CorrectRat
ER = cp.ErrorRate*
CFR = cp.Classifie
SV = cp.Sensitivit
SC = cp.Specificity
PPV = cp.PositivePredictiveValue
```

Figure.3. Data Classification using Support Vector Machine method for the given chunk size 80 in glass dataset.

V. RESULT ANALYSIS

Comparison table for the SVM and MSVM-ACO method on the basis of given different chunk size and find the value of CR, ER, PPV, NPV, PL and NL.

Table.1. Show that the calculated result of feature selection on different chunk size on the basis of two methods SVM and MSVM-ACO for glass data set.



CHUNK SIZE	SVM						MSVM-ACO					
	CR	ER	PPV	NPV	PL	NL	CR	ER	PPV	NPV	PL	NL
80	51.28	48.72	0.221	1	1.63	0	82.05	17.94	1	0.82	1.77	0.8
90	56.8	43.18	0.406	1	1.56	0	70.45	29.54	0.5	0.73	2.38	0.85
100	63.26	36.73	0.41	1	1.22	0	65.31	34.69	0.57	0.67	2.29	0.86
110	53.7	46.3	0.47	1	1.24	0	55.56	44.44	0.45	0.58	1.12	0.97

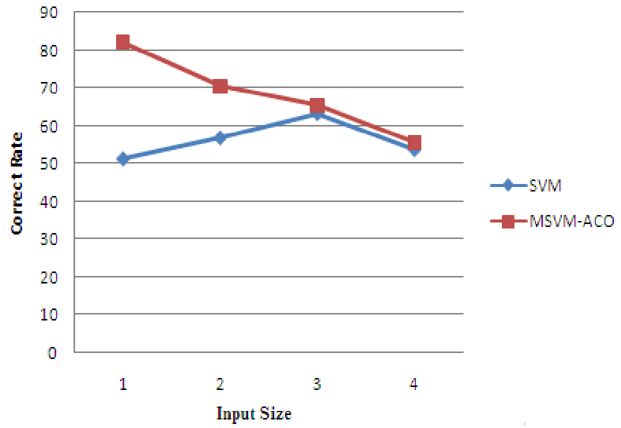


Figure.4. comparison between methods SVM and MSVM-ACO for glass data set

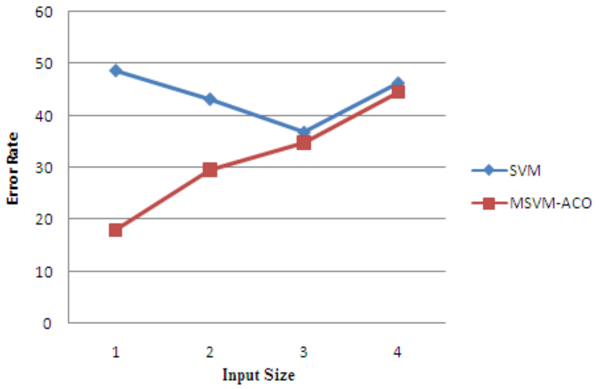


Figure. 5. Graph between error rate of methods SVM and MSVM-ACO for glass data set.

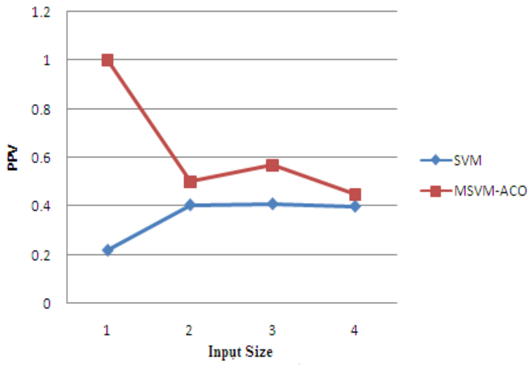


Figure .6. Graph between PPV of methods SVM and MSVM-ACO for glass data set.

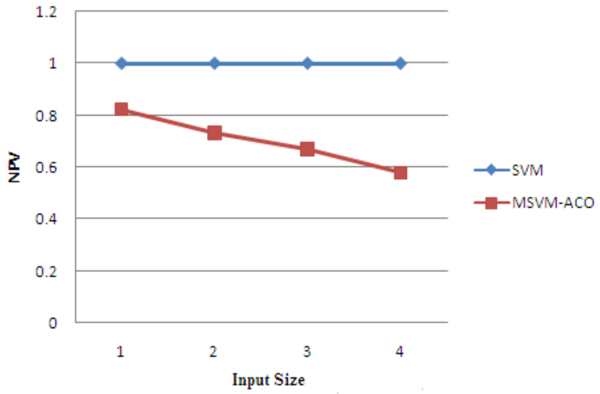


Figure.7. Shows that the comparisons graph between NPV of both methods SVM and MSVM-ACO for glass data set

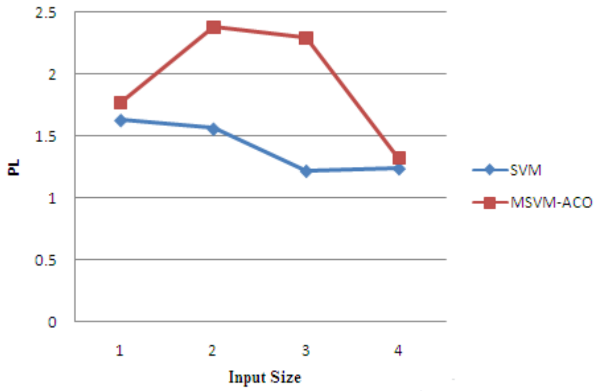


Figure .8. Graph between PL of methods SVM and MSVM-ACO for glass data set

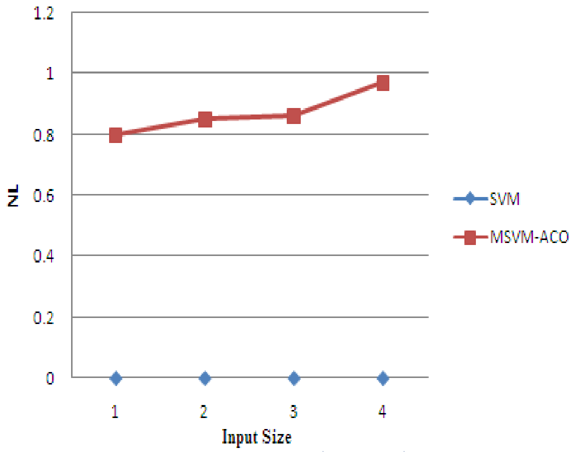


Figure .9. Graph between NL of methods SVM and MSVM-ACO for glass data set

VI. CONCLUSIONS

Traditional stream classification techniques also make impractical assumptions about the availability of labeled data. Most techniques assume that the true label of a data point can be accessed as soon as it has been classified by the classification model. Thus, according to their postulation, the existing model can be updated without delay using the labeled instance. In reality, one would not be so lucky in obtaining the label of a data instance immediately, since manual labeling of data is time consuming and costly. We claim two major contributions in novel class detection for data streams. First, we propose a dynamic selection of boundary for outlier detection by allowing a slack space outer the decision boundary. This space is restricted by a threshold, and the threshold is modified all the time to reduce the risk of false alarms and missed novel classes. Modified Support Vector Machine is very efficient data mining tool for data classification. Data classification is challenging task in the field of classification. Evaluation of new feature creates a problem in feature selection during the classification process of multiple support vector machines. In this dissertation reduces these problems using ant colony optimization, it is used to control new feature evolution problem. Ant colony optimization creates a feature prototype for cluster used in classification. The controlled feature evaluation process proposed a multiple support vector machine with ant colony optimization is called MSVM-ACO. The empirical evaluation of modified algorithm is better in comparison of SVM algorithm. The error rate of modified algorithm decreases in comparison of SVM algorithm. Also improved the rate of Correct, PPV and NPV for evolution of result, after these improvement still some problem is still remain such as infinite length and data drift. Infinite length and data drift problem are not considered in this dissertation.

VII. SCOPE OF FUTURE WORK

The proposed method multiple support vector machine solved the problem of feature evaluation and concept evaluation. The controlled feature evaluation process increases the value of correct rate and reduces the error rate. The ant colony optimization cluster faced a problem of right number of cluster, in future used self optimal clustering technique along with other optimization technique is as particle of swarm optimization, continuous orthogonal ant colony optimization etc.

VIII. REFERENCES

[1]. Heling Jiang, An Yang, Fengyun Yan1 and Hong Miao, “Research on Pattern Analysis and Data Classification Methodology for Data Mining and Knowledge Discovery”, International Journal of Hybrid Information Technology Vol.9, No.3 (2016), pp. 179-188, 2016.

[2]. Risto Vaarandi and Mauno Pihelgas, LogCluster - “A Data Clustering and Pattern Mining Algorithm for Event Logs”, International Conference on Network and Service Management, CNSM, 2015.

[3]. Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding, “Data Mining with BIG DATA”, IEEE Tran. on Knowledge and Data Engineering, 2014.

[4]. Mohammad M. Masud, Qing Chen, Latifur Khan, Charu C. Aggarwal, Jing Gao, Jiawei Han, Ashok Srivastava “Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams” TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE 2012. Pp 1-14.

[5]. Chang-Dong Wang, Jian Huang La, Dong Huang, Dong Huang “SVS tream: A Support Vector-Based Algorithm for Clustering Data Streams” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, 2013. Pp 1410-1425.

[6]. Xiao-Li Li, Philip S. Yu, Bing Liu, See Kiong Ng “Positive Unlabeled Learning for Data Stream Classification” SIAM, 2010. Pp 259-270.

[7]. Mohammad M. Masud, QingChen, Jing Gao, Latifur Khan, JiaweiHan, Bhavani Thuraisingham “Classification and Novel Class Detection of Data Streams in a Dynamic Feature Space” ECML PKDD, 2010. Pp 337-352.

[8] Huan Liu, Hiroshi Motoda, Rudy Setiono, Zheng Zhao “Feature Selection: An Ever Evolving Frontier in Data Mining” JMLR: Workshop and Conference Proceedings, 2010. Pp 4-13.

[9]. Xin Xu, Wei Wang, Guilin Zhang, Yongsheng Yu “An Adaptive Feature Selection Method for Multi-class Classification” 2010. Pp 225-230.

[10]. Kalyan Veeramachaneni, Weizhong Yan, Kai Goebel, Lisa Osadciw “Improving Classifier Fusion Using Particle Swarm optimization” Proceedings of the IEEE Symposium on Computational Intelligence in Multi criteria Decision Making, 2007. Pp 128-136.

[11]. Tahseen Al-Khateeb, Mohammad M. Masud, Latifur Khan, Charu Aggarwal, Jiawei Han, Bhavani Thuraisingham “Stream Classification with Recurring and Novel Class Detection using Class-Based Ensemble” 2012. Pp 1-10.

[12]. Albert Bifet, Geoff Holmes, Bernhard Pfahringer, Richard Kirkby, Ricard Gavalda “New Ensemble Methods For Evolving Data Streams” 2010. Pp 1-9.

[13]. Peng Zhang, Xingquan, Zhu LiGuo “Mining Data Streams with Labeled and Unlabeled Training Examples” Ninth IEEE International Conference on Data Mining, 2009. Pp 627-636.

[14]. Ashfaqur Rahman , Brijesh Verma “Novel Layered Clustering-Based Approach for Generating Ensemble of Classifiers” IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 22, 2011. Pp 781-792.

[15]. Mohammad M. Masud, Qing Chen, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham “Classification and Novel Class Detection of Data Streams in a Dynamic Feature Space” in ISMIS 2009, LNAI 5722, Pp 552-558.