Modern Applied Science; Vol. 10, No. 9; 2016
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

One Method to Reduce Data Classification Using Weighting Technique in SVM +

Arash Ghorban Niya Delavar[1] & Zahra Jafari[2]

[1] Faculty Member of Department of Computer Engineering and Information Technology of PNU, Iran

[2] Master of Computer Engineering of PNU, Iran

Correspondence: Arash Ghorban Niya Delavar, Faculty Member of Department of Computer Engineering and Information Technology of PNU, Iran. E-mail: a_ghorbannia@pnu.ac.ir/zj.nov1985@yahoo.com

Received: May 27, 2016 Accepted: June 8, 2016 Online Published: July 21, 2016

doi:10.5539/mas.v10n9p245 URL: http://dx.doi.org/10.5539/mas.v10n9p245

Abstract

SVM, a learning algorithm to analyze data and recognize patterns is used. But there is an important issue, replicate data as well as its real-time processing has not been correctly calculated. For this reason, in this paper we have provided a method DCSVM+ to reduce data classification using weighting technique in SVM +. The proposed method with regard to the parameters to SVM + has the optimum response time. By observing the parameter of data volume and their density, we abled to classify the size of interval as case that this classification to investigated case study reduces the running time of algorithm SVM +. Also by providing objective function of the proposed method, we abled to reduce replicate data to SVM + by integrating parameters and data classification and finally we provided threshold detector (TD) for method of DCSVM + to with respect to the competency function, we reduce the processing time as well as increase data processing speed. Finally proposed algorithm with weighting technique of function to SVM + is optimized in terms of efficiency.

Keywords: DCSVM+, Data Mining, SVM (SUPPORT VECTOR MACHINE), Data Classification, Replicate Data, Density, Threshold Detector (TD)

1. Introduction

In 2000s (AD), the use of SVM + algorithm spread as a method of Support Vector Machine (SVM) algorithm, compared to other methods such as Weighted SVM (WSVM), Multi-weight vector projection SVM (MVSVM), Generalized- Eigenvalue based Multi surface SVM (GEPSVM), has better performance.

The concept of data mining is to extract hidden information and relationships identified in large volumes of data in one or more big database with multiple algorithms. SVM is a learning algorithm related to analyze data and recognize patterns used to classify with regard to all data that is not limit and is expandable. Training using specific data is one of the methods of machine learning which aims to promote better classification using data that is only available during the training phase.

SVM is the applied version and the more general state of SVM, which rely on link between systematic data and indicator, which may be at least in dealing with some similar administrative problems have the practical use of reasonable information. SVM + is an implementation based on SVM from LUPI that is rarely investigated. SVM + method using previous knowledge are to improve the performance of algorithm and reduce the amount of required data that its confidential pattern is introduced recently by VAPINK and colleagues.

With increasing use of information database and increasing efficiency of utilization of information, it is needed to provide a method to improve the efficiency of applied algorithms in data mining that can use it to reduce replicate data and reduce redundancy.

Using specific data is one of the machine learning techniques that aims to promote better classification using data which is available only during the training phase. Several factors reduce the classification costs; including weighting factor of data that through checking input data affects the execution speed.

In this study, using weighting technique to specific data and indicator, we try to reduce data classification to be able to raise the speed of running algorithm and take better results.

2. Related Work

SVM is a learning algorithm related to analyze data and recognize considered patterns for classification by considering all data that is not limit and is expandable. Based on this algorithm, basis of algorithms have been proposed that each have their own advantages and disadvantages that try to speed up implementation and reduce the cost and improve the basic algorithm SVM.

Among algorithms that are proposed based on SVM, including: WSVM, LSSVM(Least Squares SVM), WLS SVM (Weighted Least Squares SVM), RLS SVM (Realest Squares SVM), MVSVM, GEP SVM, EMV SVM (Enhanced Multi Weight Vector Project SVM) and SVM + (Carlos, Javier, & Filiberto, 2014; Maksim, Matthias, & Bernt, 2014).

Table 1.

Problem(Section)	m	SVM+SVM	SVM+R^{sep}v.SVM	SVM+R$^{non-sep}$vs.SVM	SVM+vs.SVM+R$^{non-sep}$
Digits(3.2)	40	0.000	0.000	0.000	0.934
	50	0.002	0.023	0.000	0.472
	60	0.005	0.048	0.043	0.005
	70	0.001	0.002	0.001	0.473
	80	0.001	0.036	0.003	0.114
	90	0.000	0.005	0.002	0.398

LUPI is a machine learning model that recently proposed and to improve the classification using information that only during training is available and not at the time of the test, which is derived from training methods and human learning. LUPI goals are mimicking the behavior of the computing world in dealing with wide range supervisory trainings. This information is not available, but during training to build model helps better.

SVM + is the same applied version of LUPI (Learning Using Privileged Information). Despite the applied potential, this model in applications of LUPI and SVM + is rarely investigated for updating.

In this study, we provide a report of our efforts to conclude in the field of SVM + and research in scientific fields SVM +. Although more research in the field of theoretical and practical of LUPI and SVM + to better understanding of nature and feasibility and its limitations are required, this study aims to step forward by focus on the practical aspects of SVM +.

Conceptually, the main idea of the impact of index information is in the intensification and improvement of implementation. Investigations that conducted by two types of index information on the one hand is the real index information provided and on the other hand is the random features as removable and non-removable information. SVM + is to refer to any of the models however, SVM is used to refer to both models.

SVM has two methods in using data: in the first method, regular data and index information are used. In the second method, the regular data is used and then index information is used. The method that we have provided in this investigation, first we classify data based on the index information and then we take advantage of regular data and other inputs.

The main goal of classification is to distinguish between sets of input data that can be used as index information in LUPI but it should be noted that finding information that is used as an indicator, is not easy.

Engineers on specific problems in real-world use of SVM + then tried to optimize this algorithm using classification.

DTC SVM+ like SVM + relies on the relationship between the regular data and index information. Moreover, at least in some cases it has been shown that indicators are created randomly and may play a key role as index information. So when random indicators are considered as index information not only considered as a first choice, but are a baseline for main index information for comparison.

MVVM is an effective classifier that can investigate complex issues at the same time exclusively or with XOR. At the same time with this method, GEP SVM was proposed that is a multilevel classifier for binary classifications that is as multilevel proximal vector machines through generalized eigenvalue and in fact is an incentive to solve the problems of XOR and simultaneously reduce the computation time of SVM that instead of solving only in one level, solve it to two non-parallel level (Pechyony & Vapnik, 2010; Pechyony & Vapnik, 2011).

Still MVVM has more promising results and better generalized show from GEP SVM to carry out various categories. Use only one planning weight vector for each class to obtain a better classification is not enough so EMSVM method was provided that apply maximum distance from predicted average vector to points of different class to find better resolution, separated from MVSVM with maximum separation between classes by applying the maximum distance between average vector of different class. This issue causes EMSVM achieves more than a discriminative vector weight for each class according to its rating among the made scattering matrix class. In fact EMSVM can improve the classification accuracy by increasing the number of weight vectors, in fact, MVVM, GEP SVM and SVM EMSVM has a significant competitive advantage more than SVM in terms of computational cost.

The next provided method has been LMSVM that is used for the Least Strong Squares SVM for regression and classification with noise. LS SVM is least squares of SVM that are sensitive to outliers or noise in the data set. In this method, a strong novel of LS SVM is provided to avoid weight adjustment.

WLS SVM method was defined for how to assign appropriate weight to the training sample. Using P parameter to control errors and increase strength was caused the expression of RLS SVM.

RLS SVM is stronger than LS SVM, WLS SVM and LMSVM and its training time is shorter than all cases (Pechyony & Vapnik, 2010; Bollegala, Ishizuka, & Matsuo, 2011).

After these methods, SVM method was introduced by VAPINK and colleagues using previous knowledge to improve the performance of algorithm and reduce the amount of data required. SVM + is a method based on SVM using LUPI and in other words is the applied version of SVM, which relies on link between systematic data and indicator and is comparable with WSVM that based on weighting to points based on proximity or remoteness of them. SVM + executable form is on a defined set of data (Fig. 1). SVM + method provided a unique non-trivial solution much stronger than WSVM in places that wasn't offset of unit (Maksim et al., 2014).

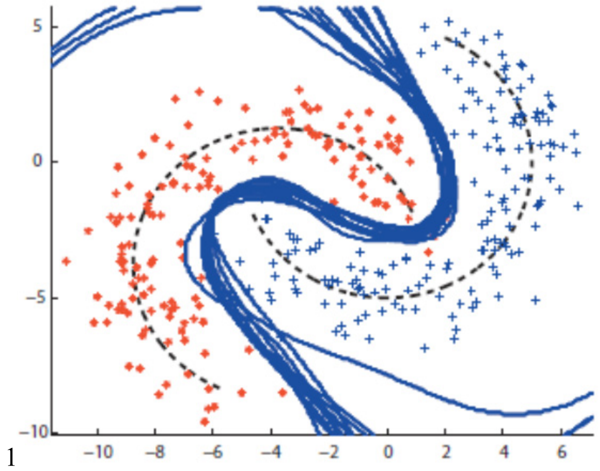

Figure 1. (Carlos et al., 2014)

By formulating SVM, the problem of duality was optimized. It should be noted that the method WSVM with regard to certain data can mimic the behavior of SVM +. In other words, any solution of SVM + is a WSVM solution by selecting the appropriate weight. But it is essential to note that any non-trivial solution of SVM + is unique, in contrast to other solutions that may be non- unique so WSVM and SVM + are not equivalent (Carlos et al., 2014; Maksim et al., 2014; Feyereisl & Aickelin, 2012).

3. The Proposed Method

Training using specific data is one of the machine learning methods that its aim is to promote better classification using data that is only available during the training phase.

In this paper we report the results of our efforts to reduce the classification of data using the weighting data.

Different methods was proposed to improve SVM algorithm that lack of appropriate generalization and lack of appropriate defined vectors caused the inefficiency of them. Finally, SVM + algorithm as a good way are known that we try to improve it in this article.

Formula SVM (Carlos Serra-Toro, 2014: 2) is defined as follows:

$$\max_{\alpha}\sum_{i=1}^{m} \alpha_I - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i\alpha_j y_i y_j K(X_i,X_j),\ \text{s.t.}\sum_{i=1}^{m} y_i\alpha_i=0, \quad 0\leq\alpha_i\leq C \tag{1}$$

Independent variable K is symbol of Kernel and C is symbol of regression.

Based on the above formula, mathematical form SVM + (Carlos Serra-Toro, 2014: 2) is defined as follows:

$$\max_{\alpha}\sum_{i=1}^{m} \alpha_I - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i\alpha_j y_i y_j K(X_i,X_j) - \frac{1}{2\eta}\sum_{i.j=1}^{m}(\alpha_i+\beta_i-C)(\alpha_i+\beta_i-C)K_z(z_i,z_j) \tag{2}$$

N is extra regression, βi is Lagrange and Z is kernel in z space.

4. Innovation and Simulation Results

Because the purpose is to reduce data classification using the weighting technique, according to the number and volume of the data instead of variable C, we used Won J, the ratio of gathering data on density of space of problem and finally we found competency function with following new form:

$$\max_{\alpha}\sum_{i=1}^{m} \alpha_I - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i\alpha_j y_i y_j K(X_i,X_j) - \frac{1}{2\eta}\sum_{i.j=1}^{m}(\alpha_i+\beta_i - \frac{W}{J})(\alpha_i+\beta_i - \frac{W}{J})K_z(z_i,z_j) \tag{3}$$

W is volume and density.

The simulation result of this competency function with preventing the data out of range, a result with density of most data and a little higher speed in performance were observed (Fig. 2)

```
alldata = [];
for i = 1:N1;
if;
size(find(data(i,:)==10000),2)<1;
alldata = [alldata;
data(i,:)].
```

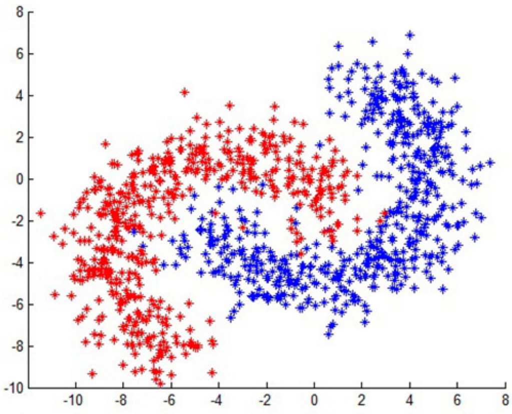

Figure 2.

Then by placing a counter, we avoided the entry of replicate data that not spend an extra time to process the data that was processed before. By doing this, speed of implementing program increases that is considered a good score in terms of time for this method (Fig. 3)

```
for j = 1:5;
testindex = [];
```

```
trainindex = [];
for k = 1:N3;
if mod(k,5)==mod(j,5);
testindex = [testindex k];
else;
trainindex = [trainindex k].
```

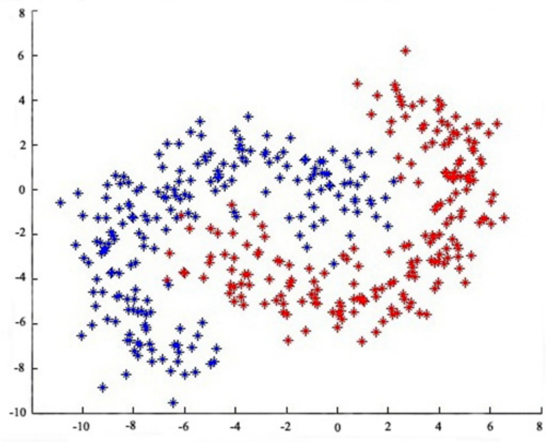

Figure 3.

Because reducing the data classification of target is considered, then by placing the threshold detector and the division of all space of data into 4 sections between zero mean and variance one with following pseudo code, this issue is improved and implemented that its results in Figures 4 to 7 are visible.

```
mdatax=mean(data(:,1));
mdatay=mean(data(:,2));
fi=(data(:,1)>=mdatax& data(:,2)>=mdatay);
datai=data(f1==1,:);
labelsi=labels(f1==1);
BEYNE 0 VA 1;
for i = 1:N2-1;
tmp1 = min(alldata(:,i));
tmp2 = max(alldata(:,i));
alldata(:,i) = (alldata(:,i)-tmp1)/(tmp2-tmp1).
```

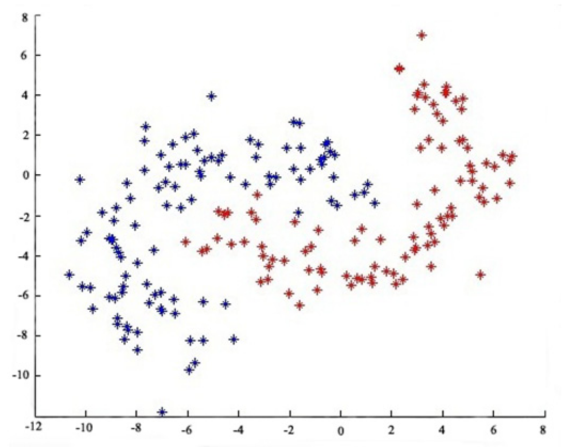

Figure 4.

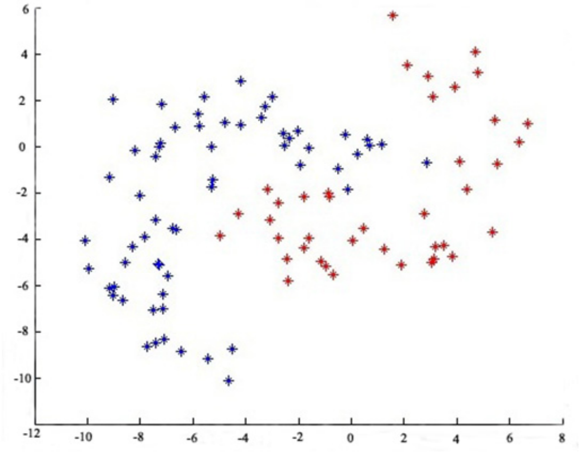

Figure 5.

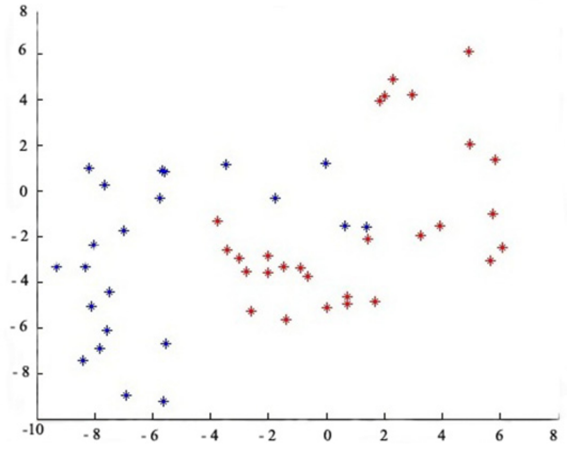

Figure 6.

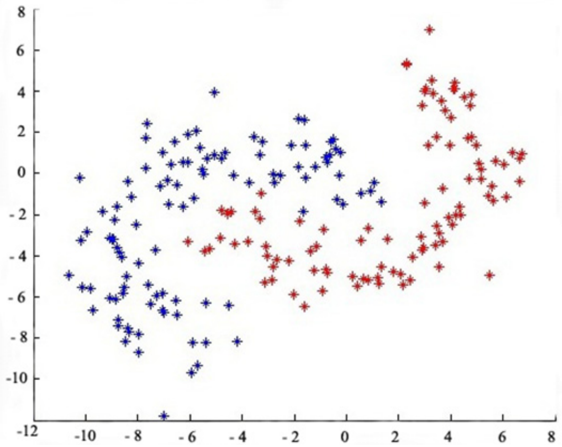

Figure 7.

5. Results

SVM have a better implementation than all SVM models. Even when use the random indicators as information. SVM + in terms of semantic and conceptual has better performance than other SVM models, so in this article was tried to improve this algorithm.

LUPI, SVM + and DCSVM + theoretically are very attractive and potentially in solving many cases are useful and effective.

If the evaluation values be replaced with a larger set of data, DSVM+ to SVM+ will be more useful in terms of computational implementation and classification.

If the error rate is increased with a reasonable difference for DSVM + and SVM +, still DSVM + has performance and better implementation than SVM +.

According to Table 1 where algorithm SVM + compared to other algorithms has been more optimum we compare the obtained results with the results of algorithm DCSVM + in the table.

Table 2.

ALGORITHM DIGIT	SVM & SVM+	SVM+ & DCSVM+
40	0/000	0/000
50	0/002	0/001
60	0/005	0/002
70	0/001	0/001
80	0/001	0/000
90	0/000	0/000

The average obtained from SVM+ algorithm is equal to 0.009 and the average obtained from DCSVM+ is equal to 0.004. It can be concluded DCSVM+ in maximum is reduced 0.005 and in minimum is equal to SVM+.

With the increase of input values in DCSVM +, evaluation forms and implementing values will be improved and has better implementation to SVM +.

References

Carlos, S. T., Javier, T. V., & Filiberto, P. (2014). Exploring some practical issues of SVM+: Is really privileged information that helps? *Pattern Recognition Letters*, 40-46. Retrieved from http://www.elsevier.com/locate/patrec

Maksim, L., Matthias, H., & Bernt, S. (2014). Learning using privileged information: SVM+ and weighted SVM. *Neural Networks, 53*, 95–108. Retrieved from http://www.elsevier.com/locate/neunet

Feyereisl, J., & Aickelin, U. (2012). Privileged information for data clustering. *Inf. Sci., 194*, 4–23.

Pechyony, D., & Vapnik, V. (2011). *Fast optimization algorithms for solving SVM+*. Statistical Learning and Data Science, Chapman and Hall/CRC.

Pechyony, D., & Vapnik, V. (2010). On the theory of learning with privileged information. *NIPS.*, 1894–1902.

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2011). A web search engine-based approachto measure semantics similarity between words. *IEEE Trans. Knowl. Data Eng., 23*(7), 977–990.

Rahman, M. M., Antani, S. K., & Thoma, G. R. (2011). Alearning-based similarity fusion and filtering approach for biomedical image retrievalusing SVM classification and relevance feedback. *IEEE Trans. Inf. Technol. Biomed., 15*(4), 640–646.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol, 2*(2011), 27:1–27:27. Retrieved from http://www.csie.ntu.edu.tw/cjlin/libsvm (accessed 19.07.13)

Ribeiro, B., Silva, C., Vieira, A., Gaspar, C. A., & Das, N. J. (2010). Financial distress model prediction using SVM+. *International Joint Conference on Neural Networks,* pp. 1–7, 2010.

Pascual, D., Pla, F., & SÙnchez, J. S. (2010). Cluster validation using information stability measures. *Pattern Recognit. Lett. 31*(6), 454–461.

Pechyony, D., Izmailov, R., Vashist, A., & Vapnik, V. (2010). *SMO-style algorithms for learning using privileged information*. Proceedings of the 2010 International Conference on Data Mining, Las Vegas, Nevada, USA, pp. 235–241, 2010.

Chang, F., Guo, C. Y., Lin, X. R., & Lu, C. J. (2010). Tree decomposition for large-scale SVM problems. *J. Mach. Learn. Res. 11*(2010), 2935–2972.

Amayri, O., & Bouguila, N. (2010). A study of spam filtering using support vector machines. *Artif. Intell. Rev. 34*(1), 73–108.

Vapnik, V., & Vashist, A. (2009). A new learning paradigm: learning using privileged information. *Neural Networks, 22*(5–6), 544–557.

Vapnik, V., Vashist, A., & Pavlovitch, N. (2008). *Learning using hidden information: master-class learning*. Proceedings of NATO Workshop on Mining Massive Data Sets for Security, IOS Press, pp. 3–14, 2008.

Liang, L., & Cherkassky, V. (2008). *Connection between SVM+ and multi-task learning*. International Joint Conference on Neural Networks, pp. 2048–2054, 2008.

Copyrights