A Comprehensive Review on Data Classification Using Support Vector Machine

Sachin[1], Deepak Kumar[2]

[1] Department of Electronics & Communication Engineering, Om institute of technology and management, Hisar- 125001, Haryana, India; sachin.svk93@gmail.com.

[2] Department of Electronics & Communication Engineering, Om institute of technology and management, Hisar- 125001, Haryana, India; ergargdeepak@gmail.com.

Abstract: In recent years, an enormous amount of research has been carried out on support vector machines (SVMs) and their application in several fields of science. SVMs are one of the most powerful and robust classification and regression algorithms in multiple fields of application. The SVM has been playing a significant role in pattern recognition which is an extensively popular and active research area among the researchers. Research in some fields where SVMs do not perform well has spurred development of other applications such as SVM for large data sets, SVM for multi classification and SVM for unbalanced data sets. Further, SVM has been integrated with other advanced methods such as evolve algorithms, to enhance the ability of classification and optimize parameters. SVM algorithms have gained recognition in research and applications in several scientific and engineering areas. This paper provides a brief introduction of SVMs, describes many applications and summarizes challenges and trends. Furthermore, limitations of SVMs will be identified. The future of SVMs will be discussed in conjunction with further applications. The applications of SVMs will be reviewed as well, especially in the some fields.

Keywords: *Classification, SVM ,Machine learning*

1. Introduction

Machine Learning is a highly interdisciplinary field which builds upon ideas from cognitive science, computer science, statistics, optimization among many other disciplines of science and mathematics. In machine learning, classification is a supervised learning approach used to analyze a given data set and to build a model that separates data into a desired and distinct number of classes [1].

There are many good classification techniques in the literature including k-nearest neighbour classifier [2], Bayesian networks [3], artificial neural networks [4], decision trees [5] and SVM [6]. K-nearest-neighbour methods have the advantage that they are easy to implement, however, they are usually quite slow if the input data set is very large. On the other hand, these are very sensitive to the presence of irrelevant parameters [7].

Decision trees have also been widely used in classification problems. These are usually faster than neural networks in the training phase, however, they do not have flexibility to modelling the parameters. Neuronal networks are one of the most used techniques. Neural networks have been widely used in a large number of applications as a universal approach. However, many factors must be taken into account to building a neural net- work to solve a given problem: the learning algorithm, the architecture, the number of neurons per layer, the number of layers, the representation of the data and much more. In addition, these are very sensitive to the presence of noise in the training data [8].

From these techniques, SVM is one of the best known techniques to optimize the expected solution [9]. SVM was introduced by Vapnik as a kernel based machine learning model for classification and regression task. The extraordinary generalization capability of SVM, along with its optimal solution and its discriminative power, has attracted the attention of data mining, pattern recognition and machine learning communities in the last years. SVM has been used as a powerful tool for solving practical binary classification problems. It has been shown that SVMs are superior to other supervised learning methods [10]. Due to its good theoretical foundations and good generalization capacity, in recent years, SVMs have become one of the most used classification methods.

Decision functions are determined directly from the training data by using SVM in such a way that the existing separation (margin) between the decision borders is maximized in a highly dimensional space called the feature space. This classification strategy minimizes the classification errors of the training data and obtains a better generalization ability, i.e., classification skills of SVMs and other techniques differ significantly, especially when the number of input data is small. SVMs are a powerful technique used in data classification and regression analysis. A notable advantage of SVMs lies in the fact that they obtain a subset of sup- port vectors during the learning phase, which is often only a small part of the original data set. This set of support vectors represents a given classification task and is formed by a small data set[11].

The rest of this paper is divided as follows: in Section 2 the theoretical basis of SVM are presented; in addition, their characteristics, advantages and disadvantages are described. In Section 3 weaknesses of SVM are introduced and reviewed. In Section 4 a set of SVM implementations are presented. Section 5 shows some applications of SVM in real world problems. Finally, Section 6 closes the paper with trends and challenges[12].

2. Theoretical basis of SVMs

The principal objective in pattern classification is to get a model which maximizes the performance for the training data. Conventional training methods determine the models in such a way that each input–output pair is correctly classified within the class to which it belongs. However, if the classifier is too fit for the training data, the model begins to memorize training data rather than learning to generalize, degrading the generalization ability of the classifier[13].

The main motivation of SVM is to separate several classes in the training set with a surface that maximizes the margin between them. In other words, SVM allows to maximizing the generalization ability of a model[14]. This is the objective of the Structural Risk Minimization principle (SRM) that allows the minimization of a bound on the generalization error of a model, instead of minimizing the mean squared error on the set of training data, which is the philosophy often used by the methods of empirical risk minimization.

In this Section, we discuss Support Vector Machines, in which training data are linearly separable in the input space and the case where training data are not linearly separable.

3. Weaknesses of SVM

Despite the generalization capacity and many advantages of the SVM, they have some very marked weaknesses, among which are: the selection of parameters, algorithmic complexity that affects the training time of the classifier in large data sets, development of optimal classifiers for multi-class problems and the performance of SVMs in unbalanced data sets.

3.1. Algorithmic complexity

Maybe the principal disadvantage of SVM is due to its excessive computational cost in large data sets, because the training kernel matrix grows in quadratic form with the size of the data set, which provokes that training of SVM on large data sets is a very slow process. Support Vector Machines (SVM) have demonstrated highly competitive performance in many real-world applications. How- ever, despite its good theoretical foundations and generalization performance, SVM is not suitable for large data set classification.

Training an SVM is usually posed as a quadratic programming (QP) problem to find a separation hyperplane which implicates a matrix of density n * n , where the n is the number of points in the data set. This needs huge quantities of computational time and memory for large data sets, so the training complexity of SVM is highly dependent on the size of a data set [15].

According to the strategy used, the training methods for SVM can be categorized into data selection, decomposition, geometric, parallel implementations and heuristics. Their core ideas and the most representative algorithms are presented in this section.

Data selection methods for SVM intent to decrease the size of data sets by removing the instances that do not contribute to the definition of the optimal separating hyperplane. The latter depends completely on instances which are located closest to the separation boundary [16], and correspond to those whose Lagrange multipliers are greater than zero in the Karush–Kuhn–Tucker conditions.

4. Performance of SVMs in imbalanced datasets

In imbalanced data sets, the correct classification of minority class objects is a challenging problem. Normal classification methods, such as support vector machines, do not work well for these skewed data sets because is difficult to get the optimal separation hyperplane for an SVM trained with imbalanced data.

The imbalance in data sets affects considerably the performance of most classifiers. In general, the model extracted from this type of data sets is biased towards the minority class. As a result, the accuracy on the minority classes is hampered. The imbalance in data sets is a recurrent problem in many domains, some examples are: fraud detection problems [17], classification of protein sequences [18], medical diagnosis of rare and dangerous dis- eases, intrusion detection and text classification, discrimination between earthquakes and nuclear explosions. Support Vector Machines were introduced by Vapnik [19] as a kernel based machine learning model for classification and regression tasks. The generalization capabilities and discriminative power of SVM have attracted the attention of practitioners and theorists in last years. SVM has strong theoretical foundations, and, in general, it presents high classification accuracy in real- world applications. However, recent experiments show that the performance of SVM is severely affected when it is applied on imbalanced data sets. This is more evident when the ratio between the majority and the minority class is large. The first dis- advantage of SVM on imbalanced data sets is due to the margin obtained is biased towards the minority class.

There are several solutions of SVM classification for imbalanced data [20]. The techniques used to minimize the negative effect of imbalanced data sets on classifiers can be categorized as external and internal. The first techniques balance the data sets before training a classifier [21]. The second techniques modify the model or architecture of classification methods. Principal external techniques are under sampling and over sampling. In general, under sampling consists in selecting, randomly, a small number of objects from majority class [146]. Over sampling techniques generate artificial examples of the minority class. Other methods use evolutionary algorithms to balance the data sets [22]. However, to add artificial data points to the minority class is a promising technique to tackle the problem of imbalance. Chawla et al. [23] proposed Synthetic Minority Over sampling Technique (SMOTE), which generates artificial objects to be included as members of the minority class. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any or all of the k minority class nearest neighbours. It does not cause any information loss and could potentially find hidden minority regions. The disadvantage of this method is that it creates noise for the classifiers which could result in a loss of performance because SMOTE makes the assumption that the instance between a positive class instance and its nearest neighbours is also positive.

4. SVM implementations

Currently there are several implementations of SVM in the literature. SVMs must solve a quadratic programming problem to find a hyperplane that separates the classes. The main reason for multiple implementations is because computational time depends mainly on the heuristics used to divide the problem into small fragments. In small data sets, the computational time of the SVMs is not important, however the computational complexity of the SVMs is almost cubic, so that in large data sets the training time is enormous and it is very important to use some algorithm that face this challenge. This section briefly shows some approaches used to improve the training time of SVM[24].

Data reduction: In most cases the SVM solution is given by a small subset of data called support vectors and not by the entire data set. The basic idea is to eliminate data less likely to be support vectors and preserve the data more likely to be support vectors and train an SVM with them.

Chunking: It is based on the sparsity of the SVM. In most cases the solution of the SVM is given by a small subset of data and not by the entire data set [164–167]. Moreover, an a_i point can only be optimal if it fully satisfies the conditions of KKT. The algorithm starts selecting an arbitrary subset of the data called chunk.

The quadratic optimization problem is solved on this small ‘‘chunk” and the next chunk is obtained with the resulting support vectors and the points violating the KKT conditions[25]. The process is stopped until all the training data are considered and the chunk get all the SV. This algorithm reduces the complexity of SVM by reducing the large problem to a sequence of smaller optimization problems, iteratively determining the support vectors.

Decomposition: These methods are similar to chunking methods. However, in decomposition methods the size of the sub problems is fixed. Decomposition methods were designed to reduce the complexity to computing the full kernel matrix by solving a sequence of smaller quadratic programming sub problems. Decomposition methods tackle the problem of training an SVM by optimizing iteratively only on the variables belonging to a sub- set of tractable size. This is the so-called working or active set. The variables that do not belong to the working set are fixed and form the so-called fixed set. Decomposition methods can be classified into primal and dual methods. They aim for dual(primal) feasibility, while maintaining primal (dual) feasibility and complementary slackness.

A clear advantage in this scheme, in addition to its proved con- vergence [26], is that its memory requirements grow linearly with the number of training examples. On the other hand, because only a fraction of the variables is being considered in each iteration, it is time consuming if elements in the working set are not carefully selected. It has been observed that the active set method can oscillate nearby the solution [27].

The most important element in decomposition methods for them to converge quickly is the selection of the subset of variables n the working set [28]. One method, commonly used, consists in selecting those samples that violated the most KTT conditions.

Sequential Minimal Optimization: The Sequential Minimal Optimization algorithm (SMO) is obtained from the idea of the decomposition method to the extreme, by optimizing a mini- mum subset of only two points in each iteration[29]. The power of this technique lies in the fact that the two-point optimization problem admits an analytical solution, eliminating the need to use an iterative quadratic programming optimizer as part of the algorithm [30].

5. Trends and challenges

Large amounts of data are generated and collected at each moment. The supervised and unsupervised learning methods of machine learning are the responsible for transforming these data into useful information[31]. SVMs have proven to be one of the best supervised learning methods in various applications; however, since SVM development several challenging problems have been identified to be able to use this classifier with very large data sets, also in dynamic environments such as data streams with concept drift, in multi-class problems, in data sets with few tagged data, and the selection of the right kernel and adjusting its parameters efficiently. These challenges are more difficult to solve when two or more of them are presented simultaneously[32]. In the following, we explain in brief some of the challenge problems of SVM classifier.

6. Conclusions

Due to its good theoretical foundations and generalization capacity among other advantages, the SVMs have been implemented in many real-world applications. SVM algorithms have been implemented in many research fields like: Text (and hyper- text) categorization, Protein fold and remote homology detection, Image classification, Bioinformatics (protein classification and cancer classification), Hand-written character recognition, Face detection, Generalized predictive control and many more. Many researchers have shown that SVMs are better than other current classification techniques. However, despite SVM has some limitations related to: parameter selection, algorithmic complexity, multiclass data sets and imbalanced data sets, SVM has been implemented in many real life classification problems due to its good theoretical foundations and generalization performance.

It is important to mention that SVM is not so popular when the data sets are very large because some SVM implementations demand huge training time or in other cases when the data sets are imbalanced, the accuracy of SVM is poor, we have presented some techniques when the data sets are imbalanced. This paper describes in detail the principal disadvantages of SVM and many algorithms implemented to face these disadvantages and cites the works of researchers who have faced these disadvantages.

References

[1] Picard, Rosalind W. "Affective Computing." Technical report 321. MIT Media Laboratory Perceptual Computing Section, April 1995.

[2] Scherer, Klaus R. "What are emotions? And how can they be measured?" Social Science Information 44, no. 4, pp. 695–729, December 2005

[3] Soleymani, Mohammad, Sander Koelstra, Ioannis Patras, and Thierry Pun. "Continuous Emotion Detection in Response to Music ideos." In IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG 2011), pp. 803–808, IEEE, March 2011. 98

[4] Russell, James A. "Culture and the categorization of emotions." Psychological Bulletin 110, no. 3, pp. 426–450, 1991.

[5] Ekman, Paul, Wallace . Friesen, Maureen O'Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E. Ricci-Bitti, Klaus Scherer, Masatoshi Tomita, and Athanase Tzavaras. "Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion." Journal of personality and social psychology 53, no. 4, pp. 712–717, October 1987.

[6] Mehrabian, Albert. "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament." Current Psychology 14, no. 4, pp. 261–292, December 1996.

[7] Fontaine, Johnny R.J., Klaus R. Scherer, Etienne B. Roesch, and Phoebe C. Ellsworth. "The World of Emotions is not Two-Dimensional." Psychological Science 18, no. 12, pp. 1050–1057, December 2007.

[8] Koelstra, Sander, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. "DEAP: A Database for Emotion Analysis Using Physiological Signals." IEEE Transactions on Affective Computing 3, no. 1, pp. 18–31, January 2012.

[9] Petrantonakis, Panagiotis C., and eontios J. Hadjileontiadis. "Emotion Recognition from Brain Signals Using Hybrid Adaptive Filtering and Higher Order Crossings Analysis." IEEE Transactions on Affective Computing 1, no. 2, pp. 81–97, August 2010.

[10] Nicolas-Alonso, Luis Fernando, and Jaime Gomez-Gil. "Brain Computer Interfaces, a Review." Sensors 12, no. 2, pp. 1211–1279, January 2012.

[11] Rosdahl, Caroline Bunker, and Mary T. Kowalski. "Textbook of basic nursing." Lippincott Williams & Wilkins, 2011.

[12] Pelvig, D.P., H. Pakkenberg, A.K. Stark, and B. Pakkenberg. "Neocortical glial cell numbers in human brains." Neurobiology of Aging 29, no. 11, pp. 1754–1762, November 2008. [13] Kandel, Eric R., James

H. Schwartz, Thomas M. Jessell, Steven A. Siegelbaum, and A. J. Hudspeth. "Principles of Neural Science." McGraw-hill New York, 2013.

[14] Khalid, Muhammad Bilal, Naveed Iqbal Rao, Intisar Rizwan-i-Haque, Sarmad Munir, and Farhan Tahir. "Towards a Brain Computer Interface Using Wavelet Transform with Averaged and Time Segmented Adapted Wavelets." In 2nd international conference on Computer, control and communication, pp. 1–4, IEEE, February 2009.

[15] Baillet, Sylvain, John C. Mosher, and Richard M. eahy. "Electromagnetic brain mapping." IEEE Signal Processing Magazine 18, no. 6, pp. 14–30, November 2001.

[16] Laureys, Steven, Melanie Boly, and Giulio Tononi. "Functional Neuroimaging." In The Neurology of Consciousness, pp. 31–42, 2009.

[17] Kübler, Andrea, Boris Kotchoubey, Jochen Kaiser, Jonathan R. Wolpaw, and Niels Birbaumer. "Brain-computer communication: unlocking the locked in." 127, no. 3, pp. 358–375, 2001.

[18] Aftanas, .I., and S.A. Golocheikine. "Human anterior and frontal midline theta and lower alpha reflect emotionally positive state and internalized attention: high-resolution EEG investigation of meditation." Neuroscience Letters 310, no. 1, pp. 57–60, September 2001.

[19] Klimesch, Wolfgang. "EEG-alpha rhythms and memory processes." International Journal of psychophysiology 26, nos. 1-3, pp. 319–340, 1997. 93

[20] Pfurtscheller, Gert, and Christa Neuper. "Motor imagery and direct brain computer communication." Proceedings of the IEEE 89, no. 7, pp. 1123–1134, July 2001.

[21] Hans, S. and Ghosh, S.(2020), "Position analysis of brushless direct current motor using robust fixed order H-infinity controller", Assembly Automation, Vol. 40 No. 2, pp. 211-218.

[22] S. Hans,S. Gupta Algorithm for Signature Verification Systems National conference on Signal & Image Processing(NCSIP-2012), Sri sai Aditya Institute Of Science & Technology.

[23] S. Hans, S. Gupta Preprocessing Algorithm for Offline signature System" National Conference on Recent Trends in Engineering & science (NCRTES- 2012), Prestige Institute of Engineering & science, Indore.

[24] S. Hans, An Algorithm for Speed Calculation of a Moving Object For visual Servoing Systems International Conference on VLSI, Communication and Networks (VCAN-2011), Institute of Engineering & Technology Alwar-2011.

[25] S. Hans & SG Ganguli (2012) Optimal adaptive Visual Servoing of Robot Manipulators

[26] S. Katyal, S. Raina and S. Hans. "A Energy Audit on Gujarat Solar Plant Charanka." International Journal for Scientific Research and Development 5.4 (2017): 2133- 2138.

[27] S. Hans (2018) A Review of Solar Energy And Energy Audit on Harsha Abacus Solar Plant: A Energy Audit on Gujarat Solar Plant Charanka.

[28] S. Hans, R. Walia ,Thapar Institute of Engineering and Technology, Patiala. "Optimal MPPT Control using Boost Converter for Power Management in PV- Diesel Remote Area ." Global Research and Development Journal For Engineering 45 2019: 24 - 31.

[29] S.Katyal, S.Raina and S. Hans. "A Brief Comparative Study of Solar Energy." International Journal for Scientific Research and Development 5.4 (2017): 2126-2132.

[30] S. Hans and S. Ghosh, "H-infinity controller based disturbance rejection in continuous stirred-tank reactor," Intelligent Automation & Soft Computing, vol. 31, no.1, pp. 29–41, 2022.

[31] S. Hans, S. Ghosh, S. Bhullar, A. Kataria, V. Karar et al., "Hybrid energy storage to control and optimize electric propulsion systems," Computers, Materials & Continua, vol. 71, no.3, pp. 6183–6200, 2022

[32] S. Hans, S. Ghosh, A. Kataria, V. Karar and S. Sharma, "Controller placement in software defined internet of things using optimization algorithm," Computers, Materials & Continua, vol. 70, no.3, pp. 5073–5089, 2022