Australian Journal of Basic and Applied Sciences
2019 December;13(12): pages 61-71
DOI: 10.22587/ajbas.2019.13.12.10
Original paper
AENSI Publications
Journal home page: www.ajbasweb.com

A Review of Parallel Support Vector Machines (PSVMs) for Big Data classification

Iatimad Satti Abd Elkarim [1], Johnson Agbinya[2]

1Sudan University of Science & Technology, Faculty of Computer Science &Information Technology, computer science department, Box 407, Khartoum, Sudan.
2School of Information Technology and Engineering, Melbourne Institute of Technology, box3000, Melbourne, Australia, jagbinya@mit.edu.au

Correspondence Author: Iatimad Mohamed Satti: Sudan University of Science & Technology, Faculty of Computer Science &Information Technology, computer science department, Box 407, Khartoum, Sudan.
E-mail:eatimadsatti@hotmail.com
Phone number:00249902664358

Received date: 15 September 2019, Accepted date: 12 December 2019, Online date: 31 December 2019

Abstract

Extending technology capability and growth of data have resulted in the need for processing large data sets faster and accurately. Machine learning techniques are used excessively to represent knowledge and classify big data. This study aims to deal with big data analysis, using parallel computing through k-means clustering applied to SVM algorithm Support Vector Machines are a reliable, efficient classification method in the area of machine learning because it has a good generalization capability and ability to classify big data accurately. However, canonical SVM is not suitable for big data sets due to its high computational complexity. Many scientists and researchers are therefore worried about how to improve the computation speeds and efficiency of different classification algorithms, and substantial accomplishments have been made.

This paper gives a review of articles of the current state of research regarding an improved form of SVM and Parallel Support Vector Machine (PSVMs) based on MapReduce and their applications in different fields. The paper further applies PSVM on realistic data. k-means clustering is used for partitioning the data points and then applied to support vector machine model. These two algorithms are implemented in four datasets for classification. Real water quality dataset from the ministry of health and different water stations in Sudan (2006-2017) is used to classify whether the water is suitable for drinking or not. The Adult dataset is used to classify the income of a person, whether it exceeds $50k/yr or not based on different parameters. The diabetes data set is used to classify whether the patient has diabetes or not based on various attributes. The Cover type dataset is used to do the classification process to five types of forest areas found in the Roosevelt National Forest in Colorado. The results of both algorithms are compared. The results showed that the applying of the k-means with the support vector machine give very strong accuracy and has a good impact on reducing computation time. The numerical experiment of applying of the k-means applied to SVM is compared with other SVM frameworks. The performance is compared using accuracy and time consuming.

Keywords: Parallel Support Vector Machine, k- Mean Clustering, Support Vector Machine, Big Data, Classification

INTRODUCTION

Knowledge by now has to be involved in computational and technological areas. The properties of large Data have widely appealed to multiple organizations, such as health care foundation, Ministries, research areas, academic fields, etc., (Vivekanandan *et al*., 2018). Big data faces many problems, especially in computational time and storage, and this is where we are. Four real datasets are applied to two models to cope with the new era of computational time problems.

Multiple classification techniques have been suggested for big data. Most of them have limitations and weaknesses, such as less performance in a big dataset, when the training set is huge, there are low run-time performance and high computation cost. Multiple research workers used classification techniques based on MapReduce to meet these limitations (Pakize *et al.*, 2014). This paper is a review article of the application and performance of PSVMs on large scale data sets and parallel computing using different algorithms. Machine learning methods are classified into three parts supervised, Unsupervised, and Semi-Supervised Learning. SVM has been known for its efficient use in many problems because of its high-performance classification ability (Birzhandi Pardis *et al*., 2019).

Citation: Iatimad Satti Abd Elkarim and Johnson Agbinya.,2019. **A Review of Parallel Support Vector Machines (PSVMs) for Big Data classification**. Australian Journal of Basic and Applied Sciences, 13(12): 61-71. DOI: 10.22587/ajbas.2019.13.12.10

The SVM is a popular machine learning technique that has shown excellent results in different application areas such as regression, power system, hydrology, power system, and medical fields. SVM can easily be used to reduce the generalization error only by maximizing the margin. (Salim Rezvani *et al*.,2019)

Many parallel implementations for SVM are proposed, but there is no precise proposed work for every application problem. Many factors affect the efficiency of the implementation of SVM, such as optimization, parallel computing kernel function, the size and dimension of the problem, hardware architecture. It is a good idea to balance the computation time and classification accuracy (Tavara and Shirin, 2019). It is necessary to use parallel computing of SVMs to improve the performance of SVM for big data. It has already demonstrated promising results for enhancing large-scale problems. The challenge due the big data is the improvement regarding computation time, accuracy, scalability, and memory issue, sowing to the immense an increasing size of real-life data requiring a reasonable choice for end-users (Tavara and Shirin, 2019).

The main objective of this paper is to review the implementation of PSVM, and K-means clustering algorithms to big data and compare the results to the sequential SVM algorithm. In this section, we describe the literature overview by defining the terms that are related to this paper. Section two describes the related work, which contains different techniques used in classification. Section three describes the methodology. While, section four describes data selection. Section five describes data preprocessing, section six experiment and results. In final section seven conclusions are provided.

1. Definition of basic concepts

1.1. Machine Learning

The Machine Learning area developed from the Artificial Intelligence fields, which are train to stimulate the ideational abilities of humans by machines (Gunnar 2004). Machine learning is an accurate technique involved with the design and improvement of algorithms that are used as an input of experimental data, such as from sensors or databases.

The supervised and unsupervised learnings are the critical two parts of machine learning. Vapnik and *et al*. (2013) had shown that machine-learning focuses on the design of algorithms that recognize intricate patterns and make predictions and creative decisions that depend on input data. The primary task in Machine Learning is the classification (Kiran and *et al*., 2013). There are challenges front of machine learning to deal with big data such as knowledge, and visions from big data to turn its potential into real value for business decision making and scientific exploration (Zhou *et al*. 2017).

1.2. Data Mining

One of the critical scientific topics is data mining, which is useful in most scientific domains. It is a helpful manner for deriving knowledge from a mass of stored raw data (Silvia *et al.*, 2012). By using various models in data mining, human errors are greatly reduced.

1.3. Classification techniques

The classification method analyzes the big data according to it is organization. There are many challenges of big data like the load and store, the form of analysis, the processing techniques (Suthaharan 2014). With supervised learning there is a problem that involves classification, and they are regarded to be instance of machine learning (Kotsiantis *et.al*., 2007). In the first step machine learning is specified a training set of rightly classified instances of datasets, then, the method is designed using this learning for prediction. Universal Classification techniques are weak when working immediately with a huge volume of data, but PSVM can express this big data. SVM is the standard technique used for classification methods. The SVM kernels are used for an exact problem that could be applied directly., and this is the most advantage of SVM, so no need for using the feature extraction process. Because of the loss of data by the feature extraction process in huge data, the use of kernel is a critical problem. The SVM results will be significantly affected when there is too much noise in the datasets. SVM is an efficient and reliable classification routine used to manage complex data (Yu *et al.*,2003).

1.4. Big Data

Big data is unstructured data that are composite to be processed in the original database systems. Because of fast-growing data, big data doesn't deal with the rule restricting the behavior of the database architectures. This data comes from multiple different sources with complications. The developing relationship in big data is growing every day (Arun and Jabasheela, 2014). The main challenges to big data are data locating, computing functions, and the environment where algorithms are applied, and the recourses are to be used. Big data generates multiple challenges for classical Machine Learning algorithms, such as scalability, flexibility, and usability, and gives a new possibility for inspiring transformative and ML solutions to label various technical challenges and create a good impact.

1.5. Support Vector Machine

Support Vector Machine is introduced in (Vladimir and Vapnik 1995; Vapnik 2013). SVM is a supervised learning algorithm that is used for classification and regression (Gunn *et al.*,1998). The goal of SVM is to find the individual hyperplane with the maximum margin that can linearly separate the classes, as shown in (Fig1). The kernel functions are used to project the training data to a feature space of a higher dimension when it is not linearly separable in the input space., in which the linear separation becomes easier. (Fig3) Shows SVM classification when it is linear separable or nonlinearly separable (Priyadarshini *et al.*,2015). Many researchers had studied and applied SVM in many practical fields. When the number of training vectors increases, their computational and storage needed increase, and this is demonstrated in different problems of practical interest out of their reach. (Rebentrost *et al.*, 2013). Support vector machine learning aims to classify data sets where the number of training

Citation: Iatimad Satti Abd Elkarim and Johnson Agbinya.,2019. **A Review of Parallel Support Vector Machines (PSVMs) for Big Data classification**. Australian Journal of Basic and Applied Sciences, 13(12): 61-71. DOI: 10.22587/ajbas.2019.13.12.10

data is small and where regular use of statistics of large numbers cannot assure an optimal solution. Two decision boundaries on the same data are shown in (Fig1) (Robinson *et al.*, 2004).

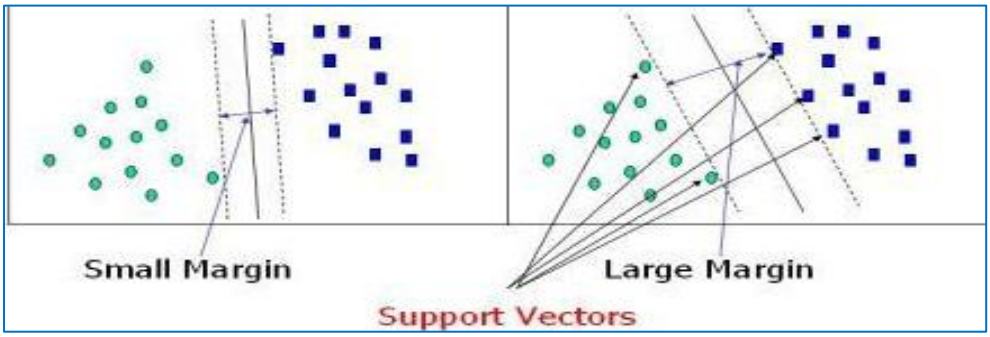


Fig1: Margin and Support Vectors. Left: small margin between 2 classes
Right: large margin between 2 classes (Nandakumar *et al.*, 2014)

In binary classification, if $((x_i, y_i) \ldots (x_n, y_n)$ are the training set, xi are vectors constitute the instances, and $y_i \in \{-1, +1\}$ are the labels of those instances. An optimum hyperplane was built by SVM, which linearly discriminates in a higher dimensional feature space that chooses the large margin that separates the two classes. The SVM classifier is shown in (Fig2). The SVM solution is to minimize the primal objective function, and this is shown in equation (1) (Ertekin *et al.*, 2011).

$$\min_{w,b} \; j(w,b) = \frac{1}{2} \| w \|^2 + c \sum_{i=1}^{n} \xi_i$$

$$with \forall_i \begin{Bmatrix} y_i (w.\Phi(x_i) = b \geq 1 - \xi_i) \\ \xi_i \geq 0, \end{Bmatrix} \quad (1)$$

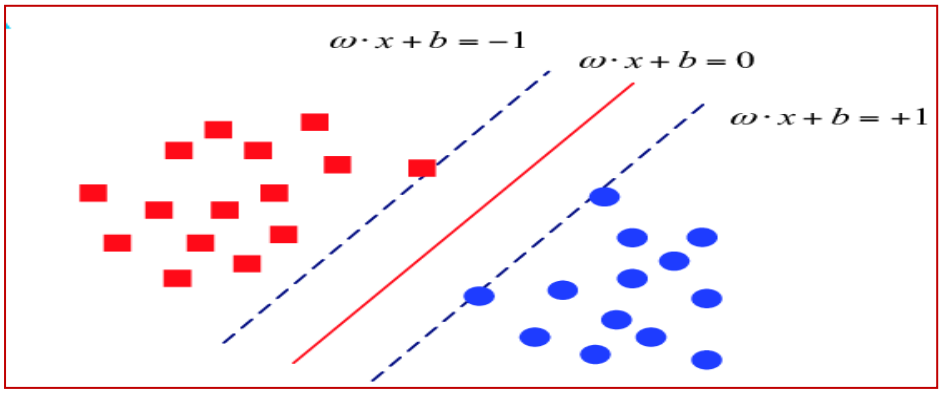


Fig 2: Support vector machine classifier

In equation (1), w is the coefficient vector of the hyperplane, b is the offset, y_i are the labels. $\Phi(x_i)$ is the mapping from input space to feature space, and ξi are the slack variables that are used with non-separable case by letting misclassification of training examples. The curved quadratic programming (QP) problem in equation (2) fixes by optimizing the dual cost function:

$$\max_{\alpha} G(\alpha) = \sum_{i=1}^{N} \alpha_i y_i - \frac{1}{2} \sum_{i.j} \alpha_i \alpha_j K(x_i . x_j).$$

$$: \quad subjectto \begin{cases} \sum_i \alpha_i \\ A_i \leq \alpha_i \leq B_i \\ A_i = \min(0, cy_i) \\ B_i = \max(0, cy_i) \end{cases} \quad (2)$$

Where $K(x_i; x_j) = (\emptyset(x_i)\emptyset(x_j))$ is the kernel matrix describing the dot products $\emptyset(x_i)$. $\emptyset(x_j)$ in feature space .

The description of traditional SVM can be as follows. Let l training samples be $T = \{(x_1, y_1), \cdots, (x_l, y_l)\}$, where $x_i \in R_n$, $y_i \in \{1, -1\}$ (classification) or $y_i \in R$ (regression), $i=1, \cdots, l$. Nonlinear mapping function is $\emptyset(x_i)$ entailing a kernel $K(x_i, x_j) = \emptyset(x_i)\ \emptyset(x_j)$. The implementation of SVM classification is solved by equation (3) (Sun and *et.al.*, 2012).

Citation: Iatimad Satti Abd Elkarim and Johnson Agbinya.,2019. **A Review of Parallel Support Vector Machines (PSVMs) for Big Data classification**. Australian Journal of Basic and Applied Sciences, 13(12): 61-71. DOI: 10.22587/ajbas.2019.13.12.10

$$\min_{w\xi b}\left\{\frac{1}{2}\|w\|^2 + c\sum_i \xi_i \right. \tag{3}$$

$$s.t.y^i(\Phi(x_i)w+) \geq 1-\xi_i \forall_i = 1,...,n \tag{4}$$

classification accuracy of the SVM algorithm is calculated as:

$$Accuracy = \frac{\text{number of support vectors}}{\text{Total data}} * 100\% \tag{5}$$

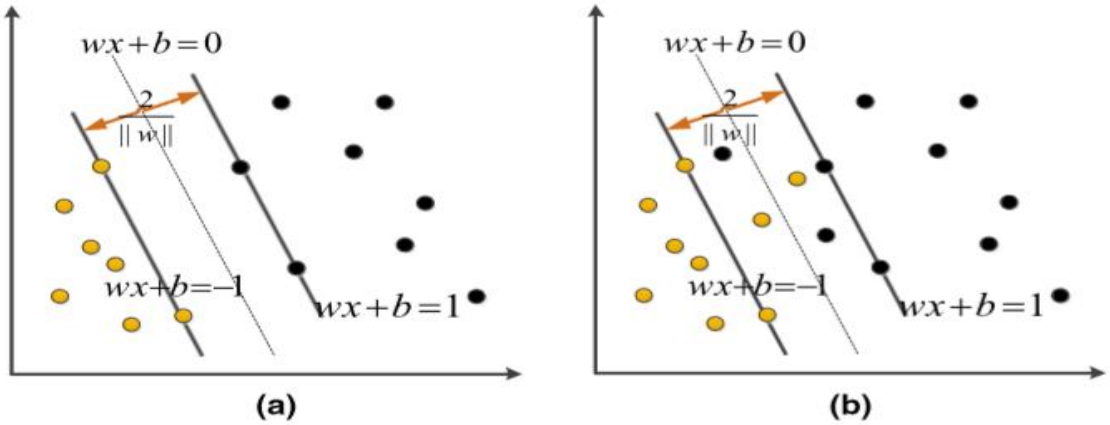


Fig3: SVM classification: (a) linear separable SVM; (b) nonlinearly separable SVM

1.6. k- Means Clustering

K-means clustering is a data mining algorithm, it is an unsupervised method that works on the resemblance. It is an iterative and efficient algorithm that works in partitioning the data points. The K-means uses one parameter (k) to specify the number of clusters, the user will determine this parameter, this clarifies that it is simple and effective. The distances between examples and clusters will be determined by using kernels. To calculate one distance, it needs to sum up all instances of a cluster (Xie *et al*., 2011; Lin *et al*., 2012). So, the number of samples and does not return a Centroid Cluster Model means that the k-means algorithm is quadratic. After the clustering is done, some clusters contain the data of two class labels called duo-cluster (Chisholm et al., 2013). The centroid determines the specific cluster in the k-means algorithm, which is the position of the center in the dimensional space of the attributes in the example Set. The k-means algorithm starts with points (k) randomly drawn from examples of the input Example Set. All Example Sets are located to their nearest cluster. Then all examples are used to recalculate the centroids of the clusters. These steps will be repeated for the new centroids until the maximum optimization steps are reached. The procedure is repeated maximum runs times (in our example maximum run =10, k=2) with different sets of start points (Chisholm *et al*., 2013).

k-mean is a heuristic algorithm, the result may depend on the initial clusters, so there is no assurance that it will be provided to the excellent global solutions. The results of the K-means algorithm are uncertain, it is be run many times, and cluster result is defined through a voting mechanism until the mean values of clusters not change. (Xie *et al*., 2011; Lin *et al*., 2012). The k- means methods steps are described as follows.

- First, the first cluster center k is chosen randomly from the database.
- Next, the first step is repeated.
- The mean value of the objects in a cluster is defined, then each object assigned to the most similar cluster.
- Then, the mean value of a cluster is updated.
- Repeat all the above steps until the mean values of clusters are not change
-

Parallel Support Vector Machine (PSVM)

When we deal with large scale datasets, regular SVM has critical issues and limitations such as complexity, size, and speed in both training and testing phase. An efficient parallel algorithm and its implementation are essential requirements to work with

Citation: Iatimad Satti Abd Elkarim and Johnson Agbinya.,2019. **A Review of Parallel Support Vector Machines (PSVMs) for Big Data classification**. Australian Journal of Basic and Applied Sciences, 13(12): 61-71. DOI: 10.22587/ajbas.2019.13.12.10

large scale data (Priyadarshini *et al*. 2015). The Cascade SVM is a series of distinct stages method that joins the results of multiple structured SVMs to create one model this is shown in (Fig4). The Cascade SVM presents several benefits over a traditional SVM because it can decrease computation time and memory size (Xiao *et al*., 2010). The main idea of the cascade SVM is to decrease a data set to its essential data points before the last step. These steps are done by locating possible support vectors and removing all other examples from the datasets, and then the collected sample datasets are preprocessed (Sagiroglu *et al*. 2013).
The cascade SVM algorithm steps:

1. First, partitioning the dataset into n subsets of similar size.
2. Next, train the SVM on each subset of the data individually.
3. Then, join the Support Vectors (SVs) of the pairs of SVMs to create new subsets.
4. Repeat, steps 2, and 3 many times.
5. At last, train the SVM on all Support Vectors (SVs) that finally received from step 4.

The cascade SVM model formed the parallel SVM. The training samples are realized through partial SVMs, the sub SVM is used as a filter, this helps to make a global optimum solution from the partial solutions. The output support vectors form sub SVM are used as the input of sub SVM in the next layer. All the sub SVM can be merged into one final SVM hierarchically (Sun *et al*., 2012). By using the PSVM model, it helps to divide the Large-scale data optimization problems into independent, smaller optimizations problems. The parallel SVM process can be shown in (Fig4).

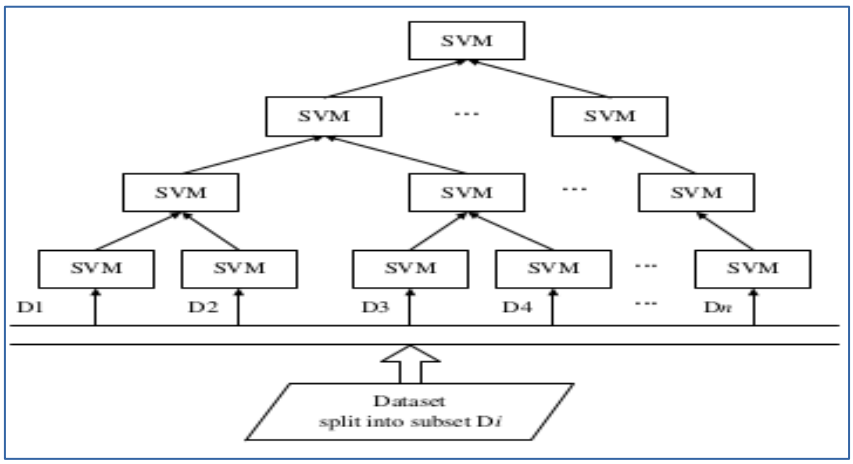


Fig4. training flow of parallel Support Vector Machine (Sun and et al. 2012).

This Cascade SVM algorithm works by runs the cascade multiple times for each data set. Next, the remaining SVs of the final model is created by combining the subsets from the first level. From the architecture of parallel SVM, we can find that it is a hierarchical structure. The low-level SVM training has to performed when all the upper-level sub SVM is trained. In the last level of the architecture, all the support vectors should be included in the training samples. The sample size must be more significant than the number of support vectors. When the ratio between the support vectors and the training sample is bigger, the speedup will be less. It is the limitation of the cascade SVM model **(**Sagiroglu *et al*., 2013).

2. Literature Review

In this section, some studies on big data methods are reviewed. These algorithms include support vector machines, parallel support vector machines (PSVMs), K-mean clustering, and other parallel computing algorithms. Big data are too large or complicated datasets that cannot be processed by regular data processing applications. The main problems with big data are its analysis, search, and it is transmitted. It requires big storage, visualization, security, and privacy of information (Nandakumar and *et al.*, 2014). There are many algorithms developed to use in the big data processing. Priyadarshini (2015), had proposed the MapReduce based on SVM used for big data and were executed on the Hadoop structure. The authors concentrated on analyzing and pruning the Kernel parameters to increase the accuracy of their research.

A model concentrating on the analysis of PSVM based on repeated MapReduce is proposed. The training examples divided into subsamples. Each sample is trained using LibSVM of the SVM model. The support vectors of any Sub-SVM counted as input to the next level. The main advantages of PSVM are reducing computation time and useful in data processing problems. (Sun et al., 2012).

An implementation method of many classification algorithms with the MapReduce model is proposed; it used in big data fields by dividing the task into small subtasks to be processed by the CPU in the cluster concurrently. (Bickson et al.,2008) had proposed a parallel execution of an SVM classifier using Message Passing Interface (MPI). They built a divided SVM solver based on the Gaussian Belief Propagation (GaBP) algorithm. GaBP is an algorithm for passing the message to apply for a message inference on graphical models (trees), showing a sustained BP where the underlying partition is Gaussian. They had enhanced the actual model by the reduction of communication load, given by the number of messages sent in every cycle, from O(n2) to O(n) aggregated messages, as n is the data points number. In the past, the GaBP algorithm was found to be so useful in parsing matrices. When we used GaBP algorithm with kernels, it will be very effective and accurate He et al. (2010).

A parallel design of SVM solver using MPI depends on a method that divided the problem into small portions of quadratic programming subproblems are shown in (Zanghirati *et al.*, 2003). These subproblems are solved using the variable

Citation: Iatimad Satti Abd Elkarim and Johnson Agbinya.,2019. **A Review of Parallel Support Vector Machines (PSVMs) for Big Data classification**. Australian Journal of Basic and Applied Sciences, 13(12): 61-71. DOI: 10.22587/ajbas.2019.13.12.10

projection method, which has proved to be accurate in solving nonlinear least-squares problems. Almost all the parameters are linear. It is efficient and quickly getting a global minimization rather than a local one. (O'leary *et al*., 2013).

The (VPM) method with a special updating rule for its projection parameter, has been studied for the Quadratic Programming problems. Then it combined the outcomes of each subproblem. The parallel execution can be used in peer-to-peer and network environments, where there is no primary authority that designates the work. Nevertheless, the drawback of support for an asynchronous connection could impact the speed of training time in the case of the implementation using the asynchronous communication model.

SVM training is mathematically a complicated process. Researchers had proposed and discovered many rules, techniques, and methods to enhance the performance of SVM (Cui *et al*., 2012). The parallel SVM model that using the Graphics Processing Unit (GPU) is proposed, by using these subsets of the training data, the training of many SVMs was done. Then the models are joined together into one model. The training data distributed into models according to their performance. The performance depends on the hardware characteristics, and the process is repeated until the rapprochement was reached. So, all tasks related to each other and have access to the kernel matrix in memory, and this leads to avoiding duplication of kernel calculations by bringing the result from memory (Li *et al*. 2013).

MapReduce method is used to disseminate the optimization problem over cloud computing structures. The proposed model of MapReduce based on parallel distribution SVM is used for binary classification. Every data set had to find the binary classifier function at its node. The algorithm collects all Support Vectors from all nodes and saves a global one (Çatak et al., 2016).

A PSVMs algorithm based on the MapReduce framework for the email classification is proposed (Xu and et al., 2014). This model was compared to the Naive Bayes (NB) model and one by one SVM, and its performance was found to be better than both of them. SVM first used to classify every email according to the field of data coming from it. Naive Bayes used for building classifiers with attributes having the highest priority. However, it has the disadvantage of independence of the predictors, which is challenging to be implemented in real life (Soni et al., 2011).

A fast parallel SVM for Large data classification is developed (Do and Nghi 2008). It used the Newton classifier to construct a parallel algorithm known as Newton's method. It reduces the quadratic approximation of the function. It is fast and effective when dealing with big data. However, it requires that the whole dataset loaded in memory. A MapReduce based SVM algorithm for MapReduce based SVM to run on files with different sizes had been developed, but the training time calculated on the Hadoop cluster. The algorithm uses graphics processors to achieve effective accuracy at low cost, and only subsets of the data are considered and loaded in memory at each time. In contrast, the solution updated in the increasing training set. The Newton SVM has extended in two ways. 1) By using GPU the developed an incremental algorithm for classifying large datasets, 2) A parallel version of the incremental Newton SVM algorithm is developed to gain high performance at a low cost. (Priyadars et al. 2015).

Zhou, Lina, et al. (2017) had introduced a framework that is applied to different problems and applications of Machine Learning (ML) on big data (MLBiD). They have presented an overview of the possibilities and difficulties of Machine Learning on big data. It consists of three phases, which are preprocessing, learning, and evaluation, and it is consists of four other components, big data, user, domain, and system. All these points and elements provide an open up future work in unknown application areas. The ML methods have not been running better while dealing with big data. So, ML and big data have to be merged to cope with the present and the future research areas.

W. Ksiaˆa et al. (2018) had proposed a mixed solution based on parallel and approx SVM for big data classification using the extended versions of the SVMs. This combination was given the name Parallel Support Vector Machines (PSVM). The main disadvantage of a PSVM model is that the feature can be deleted over time, so the accuracy is decreased. To solve this problem, they used an approach that approximates any SVM model based on the Radial Basis Function (RBF) kernel, which has been called the Approx SVM. This new approach helped to overcome two main problems, which are the failure to handle big datasets and the exchange of dimensions numbers across time. The parallel SVM has the advantage of decreasing the computation time when creating the classification model. So, the researchers in this paper had obtained spectacular results in terms of accuracy compared to the regular SVM. The parallel approx. SVM remarkably reduced the time needed to build a new model for new data (W. Ksiaˆa *et al*., 2018).

A method to define kinship relations between a presented pair of facial images using feature descriptors to study the Support Vector Machine model is proposed (Goyal *et al.* (2019). The new facial features are extracted from the salient facial features by the feature descriptors, these new features are concatenated to produce a high-dimensional feature vector. Then the SVM train these new feature vectors to classify facial images depend on similarities. A Kin Face W-I dataset is used to validate the Kinship Verification accuracy. The positive kinship pair is matched to a real parent-child pair. While negative kinship pair is matched to a pair of one's parent with another's child (Goyal et al. (2019).

3. Materials and Methods

One of the critical data mining techniques is the classification, which is used for analyzing and organizing unorganized data into an arranged class. It helps the user to gain experience and knowledge to plan for any projects. Two models are designed and compared. The first model is a single SVM trained using different kernels implementation. The second one is divided into two steps. First, the SVM is trained, then the k-means clustering is applied to this trained SVM. At last, the results of both models are compared. The framework of these models is shown in Fig5.

Citation: Iatimad Satti Abd Elkarim and Johnson Agbinya.,2019. **A Review of Parallel Support Vector Machines (PSVMs) for Big Data classification**. Australian Journal of Basic and Applied Sciences, 13(12): 61-71. DOI: 10.22587/ajbas.2019.13.12.10

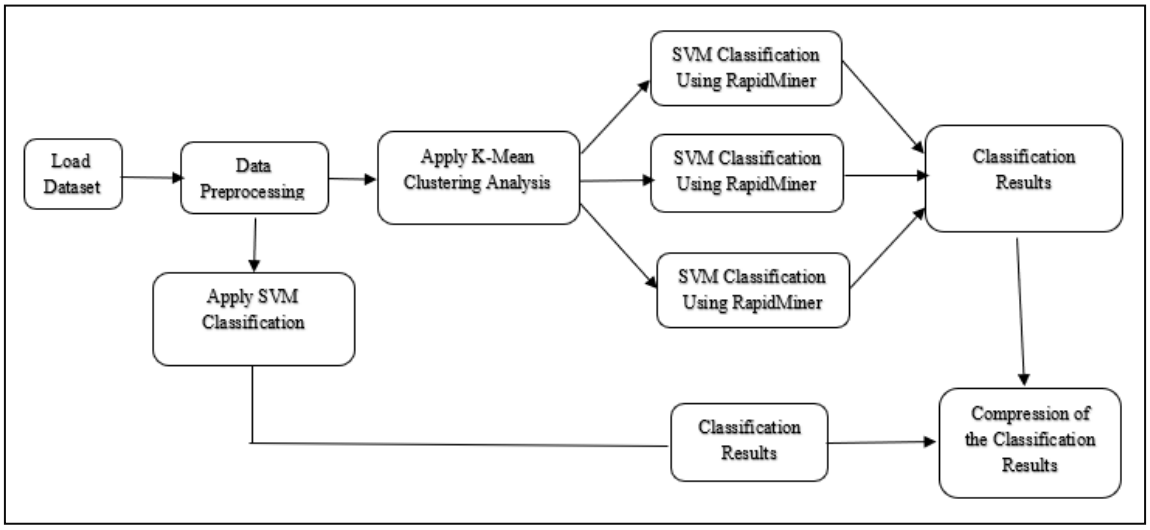


Fig 5: Methodology Framework

3.1. Dataset Selection

Four datasets are used in this paper. The first one is the Adult dataset; this dataset is extracted from the UCI repository (Lichman and Moshe 2013). It contains 42 attributes and two classes for binary classification. Attributes denoted by binary variable (0 or 1), and Labels are indicated by (+1 or -1). The second dataset is Diabetes dataset from the UCI repository. It represents ten years of clinical care at 130 US hospitals with 50 features representing the patient (Lichman and Moshe 2013). The third dataset is the River Nile water quality dataset, this dataset was obtained and provided by the ministry of health and different water stations in Directorate General of Preventive Medicine (DGPM) in Sudan from (2006-2017). It contains 20 full chemical and physical parameters. These parameters are used to predict whether water is suitable for drinking or not. The fourth dataset is the Forest Cover type datasets from the UCI repository. distinct variables are obtained from the US Geological Survey (USGS) and USFS data. (Lichman and Moshe 2013). The four datasets description is illustrated in table1.

Table 1: Datasets Description

Dataset Name	**Dataset characteristics**	**Dimensions Characteristics**	**Associated Tasks**	**Number of examples**	**Dimensions**	**Missing Values**	**Region**
Adult NEC	Multivariate	Categorical, Integer	Classification	48842	14	Yes	Social
Diabetes	Multivariate	Integer	Classification	100000	55	Yes	Life
Water quality	Multivariate	Categorical, Integer	Classification	888	20	yes	life
Cover type dataset	Multivariate	Categorical, Integer	Classification	581012	54	no	life

3.2. Data Preprocessing

Dealing with huge data is very difficult as it needs an appropriate resource and a specific environment. The four datasets shown in table1 differ in sample size and dimension. Some preprocessing is applied to datasets. Missing values are removed and replaced by the minimum value, maximum value, zero, or by calculating the average value of the specified attribute. The nominal attributes are changed to numeric attributes to cope with SVM.

3.3. Statistical Analysis

Excel 2016 is used for analyzing the results. The accuracy and execution times are used to assess the performance of the algorithms.

4. Results

This section is partitioned into two parts. The first part describes the implementation of single SVM. The second part represent the implementation of k-means clustering applied to already trained SVM.

4.1. The implementation of SVM

The four datasets are implemented using SVM algorithm. This is done by using three types of kernels, the Radial kernel has a higher accuracy for all datasets. The parameters C, gamma, and epsilon have fixed values which are, 0.0, 1.0, 0.01, respectively. Table2 and Fig6 show the results.

Citation: Iatimad Satti Abd Elkarim and Johnson Agbinya.,2019. **A Review of Parallel Support Vector Machines (PSVMs) for Big Data classification**. Australian Journal of Basic and Applied Sciences, 13(12): 61-71. DOI: 10.22587/ajbas.2019.13.12.10

Table 2. The accuracy of four deferent data sets with deferent kernel types

Kernel type	(polynomial)	(dot)	(radial)
Water quality	69.33%	69.11%	69.79%
Diabetes	67.23%	57.01%	96.40%
Adult	78.01%	76.66%	93.40%
cover type	69.90%	75.08%	74.52%

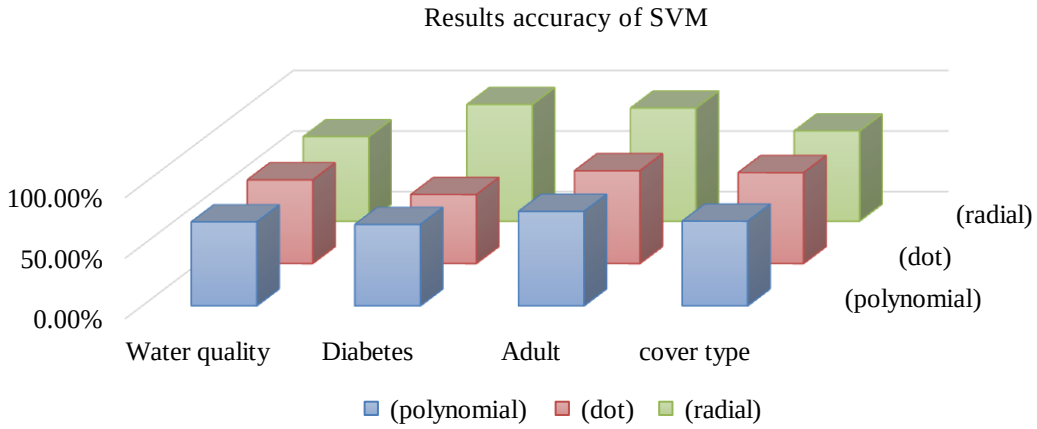


Fig 6: Results accuracy of SVM

4.2. The implementation of k-means clustering applied to SVM

The k-means clustering method is applied to SVM algorithm. The results are shown in table3 and Fig7.

Table 3: Results accuracy of SVM with k-mean (k=2)

Kernel	dot	polynomial	radial	K-mean kernel type
Water quality	70.45%	69.79%	74.52%	2
Diabetes	81.12%	76.61%	100.00%	2
adult	84.45%	87.12%	96.40%	2
Cover type	86.77%	87.12%	91.17%	2

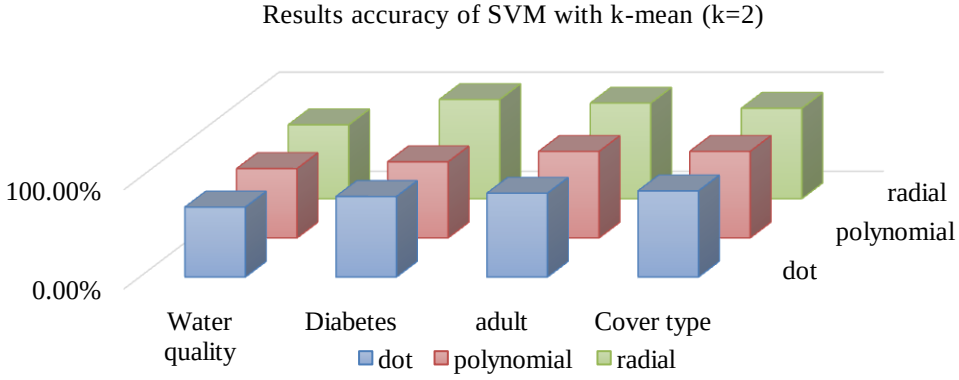


Fig 7: Results accuracy of SVM with the k-mean (k=2)

4.3. *Comparison of Both Models*

- Using the single SVM model, it is found that when the number of instances grows, the execution time increases leading to weak performance. So, k-mean clustering is used to minimize the number of these instances producing best accuracy.
- The RBF shows the best result over the other two kernels with best accuracy. For this reason, the compression between the two models is done in term of the RBF. Table4, Fig8, Fig9 show the results of this compression.

Citation: Iatimad Satti Abd Elkarim and Johnson Agbinya.,2019. **A Review of Parallel Support Vector Machines (PSVMs) for Big Data classification**. Australian Journal of Basic and Applied Sciences, 13(12): 61-71. DOI: 10.22587/ajbas.2019.13.12.10

Table 4: Result comparison

Datasets	Water quality	Adult	Diabetes	Cover type
SVM	69.79%	94.29%	96.40%	74.52%
SVM and k-means	74.52%	96.40%	100.00%	91.17%

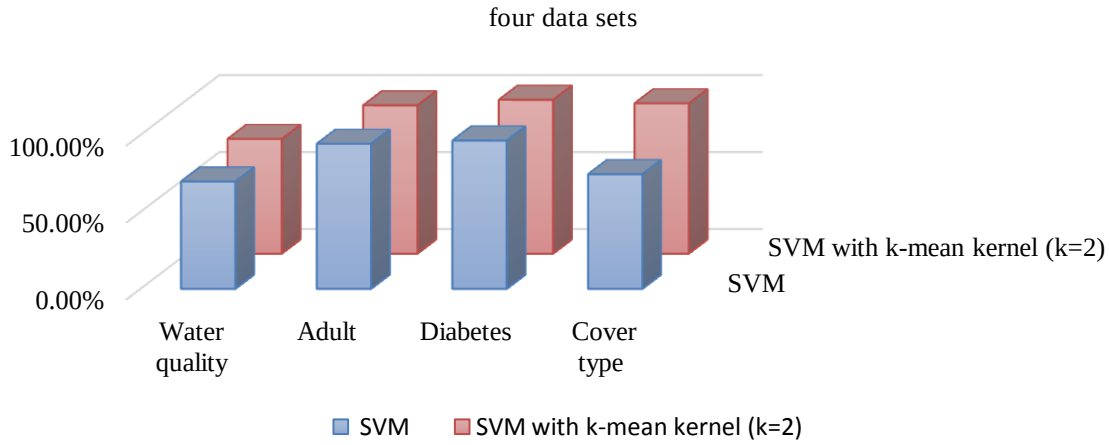
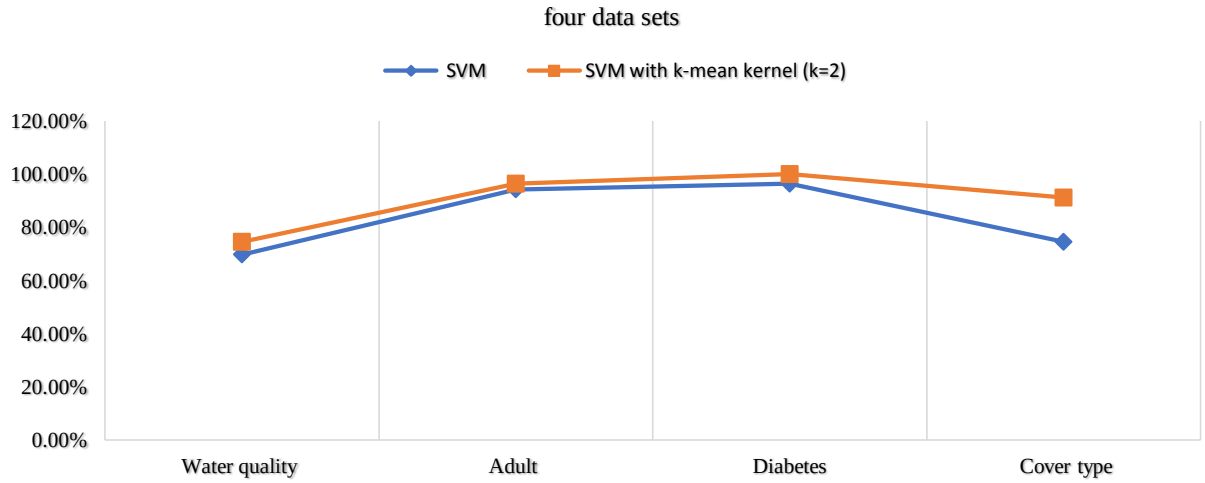


Fig 8: Results accuracy of SVM and SVM with the k-mean (k=2)



` Fig 9: Results accuracy curve of SVM and SVM with the k-mean (k=2)

5. CONCLUSION

Big data faces many problems when dealing with machine learning. Parallel computation is found to be ideal with big data, for it divides the data into smaller partitions to increase the effectiveness and accuracy and reduce the computation time. The results in this paper had shown that developing a k-means clustering model applied to an already tuned SVM algorithm produces better accuracy than traditional SVM.

It is recommended to use a parallel SVM model and compare its results to both models applied in this paper.

Acknowledgment

We are indeed grateful to our professors who help us in writing our research and document at the technical level to search methodology, results, registration, finding of the data, and discussing the results and outcome.

REFERENCES

Arun, K., and L. Jabasheela. "Big data: review, classification and analysis survey." *International Journal of Innovative Research in Information Security (IJIRIS)* 1.3 (2014): 17-23.

Bekkerman, Ron, Mikhail Bilenko, and John Langford, eds. *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press, 2011

Bickson, Danny, Elad Yom-Tov, and Danny Dolev. "A gaussian belief propagation solver for large scale support vector machines." *arXiv preprint arXiv:0810.1648* (2008)

Citation: Iatimad Satti Abd Elkarim and Johnson Agbinya.,2019. **A Review of Parallel Support Vector Machines (PSVMs) for Big Data classification**. Australian Journal of Basic and Applied Sciences, 13(12): 61-71. DOI: 10.22587/ajbas.2019.13.12.10

Birzhandi, Pardis, et al. "Reduction of Training Data Using Parallel Hyperplane for Support Vector Machine." *Applied Artificial Intelligence* 33.6 (2019): 497-516

Çatak, Ferhat Özgür, and Mehmet Erdal Balaban. "A MapReduce-based distributed SVM algorithm for binary classification." *Turkish Journal of Electrical Engineering & Computer Sciences* 24.3 (2016): 863-873.

Chisholm, Andrew. *Exploring data with RapidMiner*. Packt Publishing Ltd, 2013.

Collobert, Ronan, Samy Bengio, and Yoshua Bengio. "A parallel mixture of SVMs for very large-scale problems." *Advances in Neural Information Processing Systems*. 2002.

Collobert, Ronan, Samy Bengio, and Yoshua Bengio. "A parallel mixture of SVMs for very large-scale problems." *Advances in Neural Information Processing Systems*. 2002.

Cui, Wen, Guoyong Wang, and Ke Xu. "Parallel community mining in social network using map-reduce." *International Journal of Advancements in Computing Technology* 4.15 (2012): 445-453

Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." *Communications of the ACM* 51.1 (2008): 107-113.

Dhillon, Supreet, and Kamaljit Kaur. "Comparative Study of Classification Algorithms for Web Usage Mining." *International Journal of Advanced Research in Computer Science and Software Engineering* 4.7 (2014): 137-140.

Do, Thanh-Nghi, Van-Hoa Nguyen, and François Poulet. "A fast parallel SVM algorithm for massive classification tasks." *International Conference on Modelling, Computation and Optimization in Information Systems and Management Sciences*. Springer, Berlin, Heidelberg, 2008.

Elkano, Mikel, et al. "CHI-BD: a fuzzy rule-based classification system for big data classification problems." Fuzzy Sets and Systems 348 (2018): 75-101.

Ertekin, Seyda, Leon Bottou, and C. Lee Giles. "Nonconvex online support vector machines." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.2 (2010): 368-381.

Goyal, Aarti, and T. Meenpal. "Kinship verification from facial images using feature descriptors." *Cognitive Informatics and Soft Computing*. Springer, Singapore, 2019. 371-380.

Gunn, Steve R. "Support vector machines for classification and regression." *ISIS technical report* 14.1 (1998): 5-16.

Gunnar Ratsch, "A Brief Introduction into Machine Learning", Friedrich Miescher Laboratory of the Max Planck Society, 2004

He, Qing, et al. "Parallel implementation of classification algorithms based on MapReduce." *International Conference on Rough Sets and Knowledge Technology*. Springer, Berlin, Heidelberg, 2010

Joachims, Thorsten. "Making large-scale support vector machine learning practical, Advances in Kernel Methods." *Support vector learning* (1999)

Kiran, M., et al. "Verification and validation of MapReduce program model for parallel support vector machine algorithm on Hadoop cluster." *International Journal of Computer Science Issues (IJCSI)* 10.3 (2013): 317

Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." *Emerging artificial intelligence applications in computer engineering* 160 (2007): 3-24.

Ksiaâ, Walid, Fahmi Ben Rejab, and Kaouther Nouira. "Big Data Classification: A Combined Approach Based on Parallel and Approx SVM." International Conference on Intelligent Interactive Multimedia Systems and Services. Springer, Cham, 2018.

Li, Jie, et al. "The overview of big data storage and management." *2014 IEEE 13th International Conference on Cognitive Informatics and Cognitive Computing*. IEEE, 2014.

Li, Qi, et al. "Parallel multitask cross validation for support vector machine using GPU." *Journal of Parallel and Distributed Computing* 73.3 (2013): 293-302

Lichman, M. "UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. Irvine, CA: University of California, School of Information and Computer Science." *URL: http://archive. ics. uci. edu/ml* (2013).

Lin, Yujun, et al. "An improved clustering method based on k-means." *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*. IEEE, 2012.

Michael, Katina, and Keith W. Miller. "Big data: New opportunities and new challenges [guest editors' introduction]." *Computer* 46.6 (2013): 22-24

Nandakumar, A. N., and Nandita Yambem. "A survey on data mining algorithms on apache Hadoop platform." *International Journal of Emerging Technology and Advanced Engineering* 4.1 (2014): 563-565

O'leary, Dianne P., and Bert W. Rust. "Variable projection for nonlinear least squares problems." *Computational Optimization and Applications* 54.3 (2013): 579-593.

Pakize, Seyed Reza, and Abolfazl Gandomi. "Comparative study of classification algorithms based on MapReduce model." *International Journal of Innovative Research in Advanced Engineering (IJIRAE)* 1.7 (2014): 251-254

Perkins, Hugh, et al. "Fast parallel svm using data augmentation." *arXiv preprint arXiv:1512.07716* (2015).

Prasad, Bakshi Rohit, and Sonali Agarwal. "Handling big data stream analytics using SAMOA framework-a practical experience." *Int. J. Database Theory and Applicat* 7.4 (2014): 197-208.

Priyadarshini, Anushree. "A map reduce based support vector machine for big data classification." *International Journal of Database Theory and Application* 8.5 (2015): 77-98.

Rebentrost, P., M. Mohseni, and S. Lloyd. "Quantum support vector machine for big feature and big data classification. CoRR, vol. abs/1307.0471, 2014." (2012): 7

Rezvani, Salim, Xizhao Wang, and Farhad Pourpanah. "Intuitionistic Fuzzy Twin Support Vector Machines." *IEEE Transactions on Fuzzy Systems* (2019).

Robinson, Jonathan. *The application of support vector machines to compression of digital images*. Diss. ResearchSpace@ Auckland, 2004

Citation: Iatimad Satti Abd Elkarim and Johnson Agbinya.,2019. **A Review of Parallel Support Vector Machines (PSVMs) for Big Data classification**. Australian Journal of Basic and Applied Sciences, 13(12): 61-71. DOI: 10.22587/ajbas.2019.13.12.10

Sagiroglu, Seref, and Duygu Sinanc. "Big Data: A Review Collaboration Technologies and Systems (CTS)." *2013 International Conference on big Data*.

Silvia Sookoian, Carlos J. Pirola. (2012). The Genetic Epidemiology of Nonalcoholic Fatty Liver Disease. Elsevier Inc, 1089-3261/12, pp. 467–485

Soni, Jyoti, et al. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." *International Journal of Computer Applications* 17.8 (2011): 43-48

Sookoian, Silvia, and Carlos J. Pirola. "The genetic epidemiology of nonalcoholic fatty liver disease: toward a personalized medicine." *Clinics in liver disease* 16.3 (2012): 467-485.

Sun, Zhanquan, and Geoffrey Fox. "Study on parallel SVM based on MapReduce." *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012

Suthaharan, Shan. "Big data classification: Problems and challenges in network intrusion prediction with machine learning." *ACM SIGMETRICS Performance Evaluation Review* 41.4 (2014): 70-73.

Tavara, Shirin. "Parallel computing of support vector machines: A survey." ACM Computing Surveys (CSUR) 51.6 (2019): 123.

Thanigaivasan, Vivekanandan, Swathi J. Narayanan, and N. Ch Sriman Narayana Iyengar. "Analysis of Parallel SVM Based Classification Technique on Healthcare using Big Data Management in Cloud Storage." Recent Patents on Computer Science 11.3 (2018): 169-178

Vapnik, Vladimir. *The nature of statistical learning theory*. Springer science & business media, 2013.

Xiao, Han. "Towards parallel and distributed computing in large-scale data mining: A survey." *Technical University of Munich, Tech. Rep* (2010).

Xie, Juanying, et al. "An Efficient Global K-means Clustering Algorithm." *JCP* 6.2 (2011): 271-279.

Xu, Ke, et al. "A MapReduce based parallel SVM for email classification." *Journal of Networks* 9.6 (2014): 1640.

Yadav, Chanchal, Shuliang Wang, and Manoj Kumar. "Algorithm and approaches to handle large Data-A Survey." *arXiv preprint arXiv:1307.5437* (2013).

Yu, Hwanjo, Jiong Yang, and Jiawei Han. "Classifying large data sets using SVMs with hierarchical clusters." *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003

Zanghirati, Gaetano, and Luca Zanni. "A parallel solver for large quadratic programs in training support vector machines." *Parallel computing* 29.4 (2003): 535-551.

Zhao, Hai-xiang, and Frédéric Magoules. "Parallel support vector machines on multi-core and multiprocessor systems." *11th International Conference on Artificial Intelligence and Applications (AIA 2011)*. IASTED, 2011

Zhou, Lina, et al. "Machine learning on big data: Opportunities and challenges." *Neurocomputing* 237 (2017): 350-361