International Journal of Speech Technology (2020) 23:481–488
https://doi.org/10.1007/s10772-020-09710-1

Fusion effect of SVM in spark architecture for speech data mining in cluster structure

Jianfei Shen[1] · **Harry Haoxiang Wang**[2]

Received: 30 January 2020 / Accepted: 27 April 2020 / Published online: 12 May 2020

Abstract
Fusion effect of SVM in the Spark architecture for speech data mining in cluster structure is studied in this manuscript. Based on the information entropy of nodes, the data in clusters are fused to eliminate redundant data and improve the efficiency of information fusion. Information entropy is a statistical form based on the characteristics of information representation, which reflects the average amount of information in information. Based on the Spark platform SVM algorithm, the frequent items with the highest support after each sort are directly recursively obtained, and the transaction data set is allocated to each computing node. The structure of the item head table directly affects the efficiency of the algorithm, so optimizing the structure of the item head table can improve the efficiency of the algorithm in constructing FP-Tree, and then improve the efficiency of the whole algorithm. The proposed speech data mining algorithm can cluster, analyze, and comprehensively detection the saliency information, the detection accuracy is much higher than the state-of-the-art models. The experimental results compared with the latest research have reflected that fact that the proposed model has the better performance and robustness.

Keywords SVM · Spark architecture · Data mining · Speech analysis · Cluster structure · Fusion effect

1 Introduction

In daily communication of human beings, language communication is required, and the most important way of language communication is voice communication. The voice signal contains a lot of information, and it also has the very high variability, including changes in tone and mood, etc. Therefore, the study of this unique change can very effectively improve the success rate of artificial intelligence speech emotion recognition. In the field of artificial intelligence, in the process of human–computer communication, speech has a very high amount of information for the information transmission, fast and practical, so that the research on speech emotion recognition method has gradually become a hot research point in the near future. The technology of speech and emotion recognition is mainly through receiving the voice sent by people, and processing and judging these speech, analyzing the real intention that people want to express, so that they can carry on the human–computer interaction normally and reasonably. Although the current AI speech recognition technology has made great progress, AI's recognition of sound languages still needs to be trained and optimized with large amounts of the data according to the application scenario. Each language requires more than 10,000 h of sound data in different scenarios (Talan et al. 2019; Sreeyuktha and Reddy 2019; Maleki et al. 2019).

In the later stage, adjustment and optimization need to be performed according to the actual evaluation effect. Hence, the SVM and Spark framework should be combined to achieve the better understanding of the recognition task. Support vector machines have a strong theoretical basis. It can then ensure that the extreme value solutions found are global optimal solutions rather than local minimum values. This also determines that the SVM method has a better generalization ability for unknown samples. Advantages, SVM can be well applied to the fields of pattern recognition, probability density function estimation, time series prediction, regression estimation, etc. The regression prediction of support vector machines from linear to non-linear transformation is through a kernel function. And the kernel function is a kind of basic mapping, so if we choose different kernel

✉ Jianfei Shen
yuxiang@goperception.com

[1] Visual Arts Department, Hunan Mass Media Vocational Technical College, Changsha, Hunan, China

[2] GoPerception Laboratory, Ithaca, USA

functions, the results will form different algorithms. Moreover, the support vector machine is similar to a neural network in form, and the output is a linear combination of the intermediate nodes, and each intermediate node corresponds to a support vector. The main problem of the SVM research is the solution of the convex quadratic programming. The traditional SVM algorithm has many problems in computing, including the slow speed of the training algorithm, the complexity of the algorithm and the difficulty to realize, the large amount of operation in the test stage, the ability to resist noise and the isolation point, and so on. Therefore, in the research of SVM algorithm, how to improve the training speed, reduce the training time and establish a practical learning algorithm is an urgent problem to be solved. Therefore, Spark structure can enhance the efficiency and the robustness of the SVM analytic framework (Ray et al. 2016; Zhang et al. 2017,2018; Suthakar et al. 2016; Chen et al. 2018; Cordero et al. 2016). The Fig. 1 gives the basic framework of the Spark. In terms of data models, traditional moving target data models (such as moving target position models based on dynamic attributes, object models based on spatiotemporal geometry, and spatiotemporal data models based on events, etc.) can be used to express the spatiotemporal information of moving objects or the trajectories, but it is mainly targeted at the traditional smaller-scale traffic data management tasks. Spark Streaming can perform streaming calculations on real-time data. It is punctual and provides a high-level abstraction of discretized stream data to represent a continuous stream of data.

Spark is a low-latency cluster distributed computing system for very large data sets. By adopting the elastic distributed data set RDD, the temporary calculation data results in the MapReduce calculation process are eliminated, and a large amount of unnecessary hard disk IO overhead is effectively reduced and, Spark has inherent advantages in the real-time data processing, data mining, and machine learning, especially in calculations that require a large number of iterations. The Fig. 1 shows the framework of the Spark for parallel data processing tasks. Therefore, the Spark platform is a more versatile and lower cost data platform.

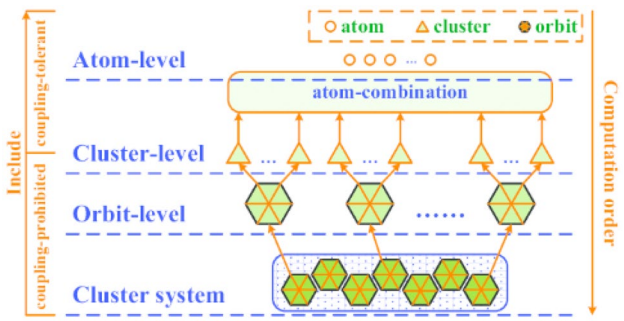

Fig. 1 The spark architecture for the parallel data processing tasks

Compared with the Hadoop platform, Spark is a memory-based computing framework and a parallel distributed computing framework based on the MapReduce and it is fast and efficient. Inspired by this, the paper will propose the fusion effect of SVM in Spark architecture for speech data mining in cluster structure. The rest of the manuscript is organized as the follows. In the Sect. 2, the Spark and data fusion model is presented. In the Sect. 3, the proposed SVM based speech analysis framework is designed. In the Sect. 4, the experiment is conducted and in the Sect. 5, the summary is arranged.

2 Spark and data fusion

2.1 The spark structure

Spark is fast general-purpose computing engine specially designed for large-scale data processing. It is a new open source platform based on the Hadoop MapReduce. Spark supports interactive computing and complex algorithms, which is fast has high versatility and also supports multiple resource managers. The protection of the data confidentiality has become the focus of attention from all walks of life (Gupta et al. (2017); Yu et al. 2016; Śmieja and Wiercioch 2017; Thakur et al. 2019; Ozcan and Basturk 2019; Song et al. 2019).

But everything is relative. Due to the rapid development of big data technology, the demand for data exchange is increasing. The issue of confidentiality of passenger data has limited data mobility. One side is the need for data protection, and the other is the need for information interconnection and interoperability in the era of big data. The Spark cluster includes a driver, a cluster manager, and various worker nodes. The driver includes a process for creating a SparkContext object to schedule and control processes. SparkContext is connected to the cluster manager. Here, Hadoop YARN cluster management is used. After the driver is connected to the cluster manager, the worker node will create a program for each application to perform tasks, including running code and data storage. After the execution program is then generated on the worker node, Spark will send the code of each application to Each executor, and SparkContext also sends tasks to each executor to run. Figure 2 gives the detail.

For effectively use the features of the Spark, the following aspects should be considered. (1) In the logistic data warehouse, the data is stored in the fact table in an exhaustive manner. Often in order to improve the query efficiency, some fact tables need to be aggregated. If MapReduce is used for aggregation, it will undoubtedly increase disk read process, resulting in higher disk overhead. In order to reduce disk I/O operations while the

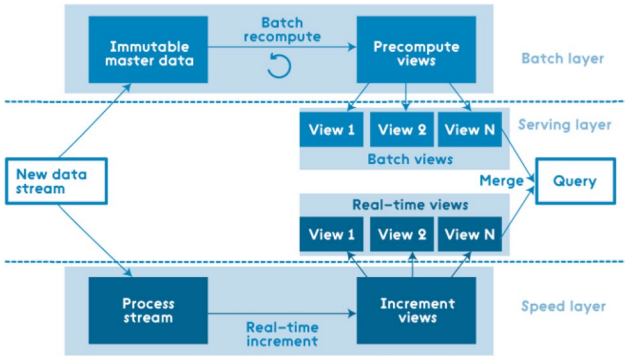

Fig. 2 The details of the spark frameworks

Spark supports direct aggregation and also supports packet aggregation. (2) After the above data conversion process, the data is in accordance with the logistics data warehouse rules, only need to load the data through HDFS logistics data warehouse to complete, this step can use the Spark SQL to improve efficiency, the data that needs to be loaded into the logistics data warehouse with Spark SQL to create the external Hive table, the address of the external Hive table to the corresponding logistics data warehouse, and then can use Spark SQL directly for efficient data query. (3) In order to deal with large data sets, random sampling and segmentation are adopted. The sampling method can reduce the data quantity and then improve the efficiency of the algorithm. Generally, better clustering results can be obtained. Segmentation means, that is, the sample is divided into several parts, and then the objects in each part are locally clustered to form sub-classes. Then cluster the sub-classes to form a new class (Souza et al. 2019; Wang et al. 2018; Lang et al. 2018; Chauhan et al. 2019; Ibrahim et al. 2019; Xiong et al. 2019).

$$P(I = c_k) = \frac{\sum_{p=1}^{N} M(i_p = c_k)}{N}, \quad k = 1, 2, ..., K \tag{1}$$

In the formula 1, we define the Spark center model. Because the CURE algorithm uses multiple objects to represent a class, and adjusts the shape of the class through the shrink factor class, it can handle non-spherical object distributions.

Elastic distributed data sets are the basic data abstractions in Spark, with features such as automatic fault tolerance, location-aware scheduling, and scalability. The emergence of Spark extends the traditional MapReduee model, which can not only provide more and more efficient computing models, but also be suitable for many different distributed platform scenarios. At the same time Spark users can simply and low-cost integration of the various processing processes, which not only meet the requirements of real-time computing and real-time processing, but also reduce the burden on the various platforms to manage separately.

The executor node is responsible for running tasks in Spark, and the executor is also started when the Spark starts. The executor also has two functions: first, it is responsible for running Spark tasks, and then returns the results to the driver; second, it provides in-memory storage for RDDs through the block manager, which can use cached data to accelerate operations.

$$P(S^{(j)} = a_{jl} | I = c_k) = \frac{\sum_{p=1}^{N} M(s_p^{(i)} = a_{jl}, i_p = c_k)}{\sum_{p=1}^{N} M(i_p = c_k)} \tag{2}$$

In the formula 2, the defined set of data is analyzed. The node of the Master corresponding to the ClusterManger process, which can control the operation of the entire cluster, is the most indispensable part of the cluster. Slave corresponds to the node of the Worker process. Its function is to report and accept specific commands and distribute specific tasks down on the present image. Exector performs the exact part of the task. Client is the interface that interacts with the user that provides the visualization and is also then responsible for the task submission. The distributed environment used is Spark + Yarn mode. Spark is used as a distributed computing framework. Yarn is used as a cluster manager. According to the cluster configuration, Spark applies 15 executors as computing resources. Each executor includes a CPU and 1000 M memory. In Spark, data is split into the several partitions, and different partitions are located in different executors. In this case, each executor only needs to calculate the data in its own partition to achieve the purpose of distributed computing.

The selection of the number of partitions is a key factor affecting the efficiency of Spark computing. Too many partitions will cause the data shuffling process to take longer, while too few partitions will make the join operation more time consuming and reduce system fault tolerance, for the better performance, data fusion is the core. There are two ways to make an image using a Docker container: (1) Use the base image to run the container, install software and perform related operations directly in the container, and then use the "docker commit" command to package the container into a new image. (2) According to the specific experimental goals and requirements, based on the original base image, configure and write the Dockerfile file, and use the "docker build" command to build a new image. Therefore, we use the latter one.

2.2 The data fusion framework

The data producer can be a human user, or a variety of data producers such as the computing devices, terminal

devices, and cameras. This is also an inevitable feature of the 5G era. The main function of data planning is to map the physical data storage partitions to logical partitions according to the needs of data users. Data source analysis, is based on the characteristics of different data sources, mining data relations and the data fusion is the key module of the system (Vapnik and Izmailov 2017; Ning et al. 2016; Deng et al. 2019; Singh et al. 2016).

Users fuse different data sources and build the multi-dimensional data models to integrate into a complete data application. Many mathematical tools and methods are used in information fusion, mainly including probability theory, fuzzy theory, wavelet method, neural network and other methods. Information fusion, as an intelligent and ultra-high data comprehensive processing technology, integrates and applies many disciplines and new technologies. The broad scope of data fusion includes detection technology, signal processing and communication technology, pattern recognition, decision theory, estimation theory, optimization theory, computer science, artificial intelligence, and neural networks (Kadyan et al. 2019; Alassaf et al. 2019; Ouahabi et al. 2019).

Hence, for the effective analysis of the model, we present the following pipelines. (1) The edge device computing power sharing management framework aggregates the unique intelligent computing capabilities possessed by general servers and intelligent terminals in the edge IoT application scenarios, and shares them with other ordinary terminals without intelligent computing capabilities by invoking network interfaces, providing them with nearby agile and secure localized intelligent computing services. (2) Through virtual description modeling, knowledge map construction and service capability abstraction of edge equipment, virtual edge equipment resource pool is formed. SDEC controller manages and controls virtual device resources through the device resource search and matching algorithm, device resource scheduling and also arrangement strategy, device resource sharing and collaboration strategy based on the knowledge reasoning. (3) The software framework uses the edge device to form the edge computing network, in which the main control node communicates upward with the cloud, and downwards communicates with the subordinate node through the collaboration engine. the subordinate node module receives the task distributed by the main control node and monitors the node running state. In the whole network, the node allocates the computing engine according to the result of the task scheduling to carry on the concrete computation task control, and also completes the data real-time processing cooperatively. The Fig. 3 gives the details.

The classical wireless communication system model is used as the core energy consumption model of the routing

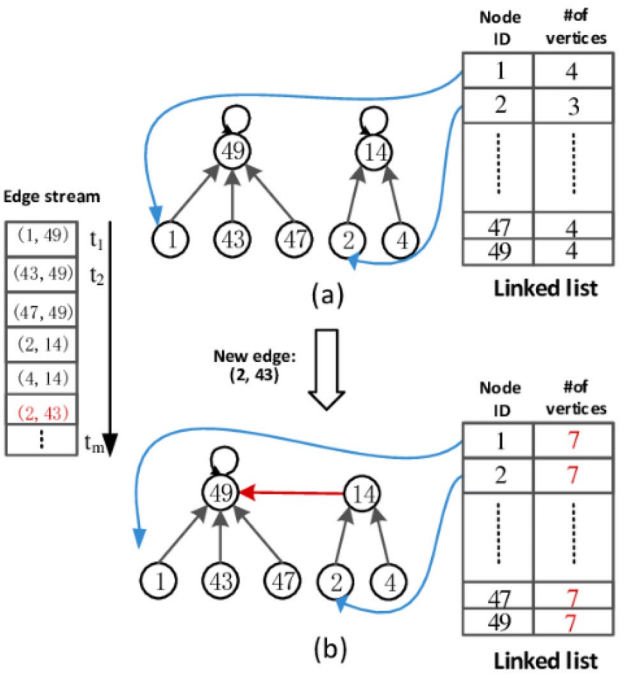

Fig. 3 The data fusion pipeline demonstration

algorithm, and the nodes in the network choose the corresponding model according to the distance they need to communicate.

$$W_{ij} = \begin{cases} \exp\left(-\left\|x_i - x_j\right\|^2 \Big/ \eta\right), & \text{if } L(x_i) \neq L(x_j) \\ 0, & \text{else} \end{cases} \tag{3}$$

In the formula 3, the analytic framework is demonstrated. The non-uniform clustering stage is divided into the election stage of cluster head and the clustering stage. Firstly, cluster heads are selected from all nodes in the region to be detected, and competition radius of candidate cluster head nodes is calculated based on the distance between nodes and base station, node density and energy, to complete cluster establishment. For improvement, the formula 4 gives the optimization.

$$\begin{aligned} J(a_i, \lambda_i) &= \sum_{i=1}^{d} a_i^T (S_b - S_w) a_i - \lambda_i (a_i^T a_i - 1) \\ \frac{\partial J(a_i, \lambda_i)}{\partial a_i} &= \left((S_b - S_w) - \lambda_i I\right) a_i = 0 \end{aligned} \tag{4}$$

Each sub-filter does not consider independent feedback and reset structure for independent filtering. This can reduce the amount of feedback information and data calculation from the main filter to the sub-filter. But if there is no effect of feedback information, the estimation accuracy

and stability of the filter will be reduced. Therefore, in our model, we will try to avoid this feature.

3 SVM based speech analysis framework

3.1 The revised SVM architecture

Support vector machine is a new machine learning method based on statistical learning theory. Machine learning mainly studies how the computers simulate or then implement human learning capabilities to the acquire new knowledge and skills, reorganize existing knowledge structures, and continuously improve performance. In the SMO algorithm whose kernel function is also dominant, the appropriate cache replacement strategy plays a very important role in the performance improvement of the algorithm, but the selection of the working set in most SMO algorithms such as Platt's heuristic, Keerthi's improvement and feasibility Direction method, the purpose is to make the objective function drop as much as possible, to find the minimum value of the objective function as fast as possible.

Based on the above analysis, by adding effective constraints and introducing Lagrangian multipliers, the optimal classification discriminant function is solved, and the determination of its parameters depends on the support vector.

In practical applications, the kernel function combined with the core optimal classification discriminant surface The support vector machine model solves the disadvantage of only processing linearly separable samples. The combination of the two forms the final support vector machine model. Support vector machine is based on two classes of the linear separable developed from sample data, but in practical applications, the need to identify and classify the data of the majority of cases are in nonlinear state, is not the ideal state. Thus, the researchers designed a kernel function is applied to the classification of support vector machine to solve the problem in the process, its main purpose is to the original nonlinear in low dimensional space cannot score according to mapped to the high-dimensional space, which can't solve the low dimensional feature space structural classification hyperplane.

The application performance of the support vector machines is the key to the selection of kernel function method. In the Fig. 4, we present the sample.

In the sample, we consider the following existing research. (1) DSVM uses dropout technology, which can select key data points while avoiding overfitting, and efficiently classify the identified objects. DSVM is composed of a stack support vector machine, which is used to automatically extract features from the original image and can use it according to its utilization. Multi-class SVM based on the RBF kernel for image classification. (2) LS-SVM

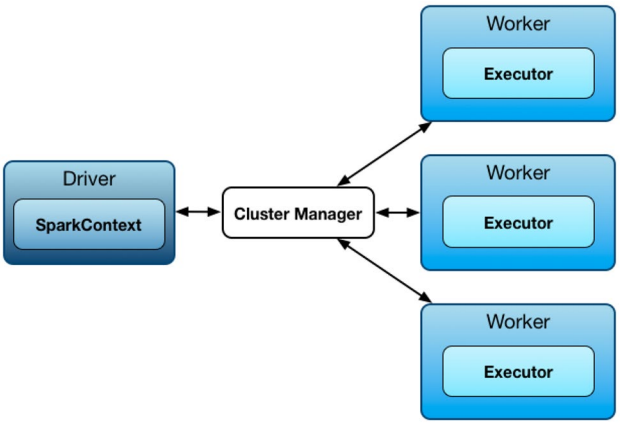

Fig. 4 The parallel SVM model

was used to establish the inverse model of the bearingless reluctance synchronous motor nonlinear system. The inverse model was connected in series with the composite object of the bearingless reluctance synchronous motor to decouple the multivariable strong coupling system into the independent subsystems. For each subsystem as an additional closed-loop controller PID was designed, and the reliability and correctness of the control method were verified by the built simulation model. (3) The inverse model of ball mill pulverizing system was established by using the SVM, the pseudo-linear composite system was constructed, and the predictive controller was designed for the closed-loop optimal control of the composite system. The effectiveness of the SVM inverse predictive control method is verified in the ball mill pulverizing system, which has good control performance in tracking characteristics, response speed and control accuracy. (4) The SVM parameters are determined by structured learning technique, and the optimal parameters of the relaxation structure SVM training estimation are also constructed. structured SVM method reduces the computational complexity of recognition and addresses the irregular and diverse interference factors of handwritten documents.

3.2 The enhanced speech analytic framework

Support vector machines can be divided into linear and nonlinear. For linear support vector machines, classification is straightforward. For a non-linear support vector machine, the data needs to be mapped into a high-dimensional space, and then it becomes linear before classification. In this paper, various characteristic parameters are extracted through speech processing, then they are combined into feature vectors, and finally emotion classification is performed. Obviously, the data here is linearly indivisible, therefore, we use nonlinear SVM for classification.

In the Fig. 5, we demonstrate the SVM model for processing the data sets. Based on this definition, this study summarizes the knowledge services of intelligent voice assistants into three levels: one is to complete the input and output of the information, that is, basic capabilities; the other is to search, organize, analyze, and reorganize information knowledge That is, primary knowledge service capabilities; the third is to provide users with personalized support and services based on user problems and environments, that is, advanced knowledge service capabilities. These three levels have become a first-level indicator of the knowledge service capability of the intelligent voice assistant (Wu et al. 2018; Sangaiah et al. 2019; Wottschel et al. 2019).

$$S_i(t) = S_i(0)\left(1 - \frac{t}{T}\right) \tag{5}$$

In the formula 5, we present the initial model adopted. Before the FastICA algorithm separation, the sound source signal needs to be pre-processed, that is, the centering and whitening. Centralization is the observation vector Z minus its mean, which becomes a zero mean vector. Generally, there is a core correlation between the obtained data variables, so the data needs to be pre-processed, that is, whitening processing. It can remove the correlation between data variables, that simplify the extraction process of independent components in subsequent processes, and improve the algorithm degree of convergence. Operation is presented as formula 6.

$$\begin{aligned} S_b &= \frac{1}{N}\sum_{j=1}^{N} W_{ij}(m_j - m_0)(m_j - m_0)^T \\ &\propto \sum W_{ij}(x_i - x_j)(x_i - x_j)^T \\ &\propto \left(XD_bX^T - XWX^T\right) \\ &= XL_bX^T \end{aligned} \tag{6}$$

FastICA algorithm is a number of ICA algorithms, because of its fast convergence speed, good separation effect is widely concerned, also known as the fixed point algorithm, is a fast optimization iterative algorithm, the Newton method used in this algorithm is the batch calculation value of multiple sampling data, that is, in each step of the iteration has a large number of sample data to participate in the calculation. For the efficient computing, the cluster pattern is then considered. With the continuous improvement of network bandwidth, service combinations such as high availability, dynamic virtual resource management, and short-term server failover that were previously only available in specific types of cluster configurations can be technically implemented and implemented in a virtual cloud computing environment with the improved performance. Keepalived is a service software that ensures high availability of the cluster in cluster management to prevent single points of failure. It is based on Virtual Routing Redundancy Protocol (VRRP). That is, the N routers that provide the same function form a router group. This group contains a master and multiple backups. The master has a VIP (providing external services). The default route of the other machines in the LAN where the router is located is this VIP. The master will send multicast. When the backup cannot receive the vrrp packet, the master is considered to be down. In this case, a backup needs to be elected according to the priority of VRRP master. In this way, high availability of the router can be guaranteed. In our designed system, we will use the VIP model to process the parallel information.

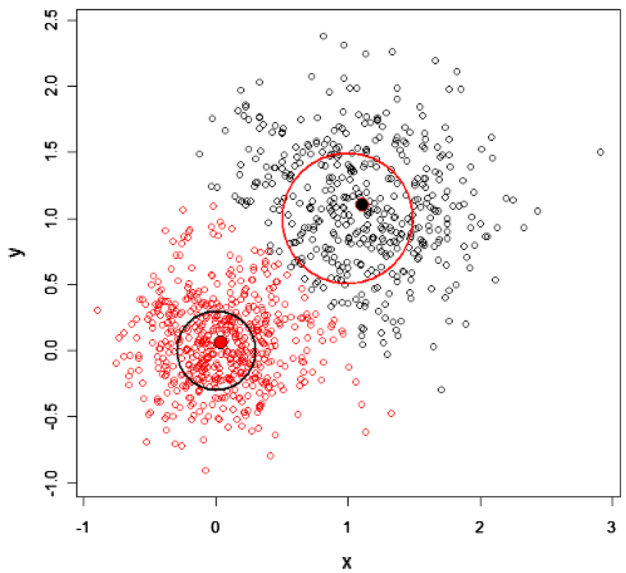

Fig. 5 The SVM for processing the data sets

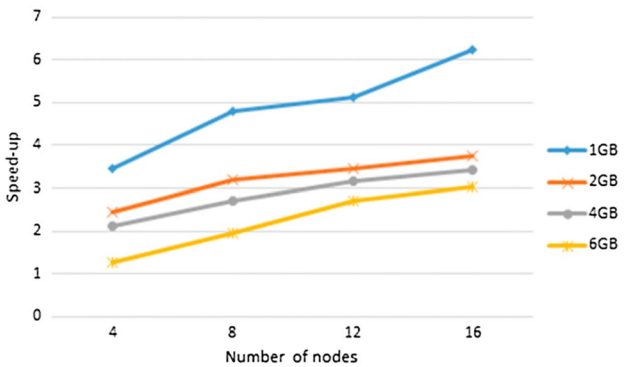

Fig. 6 The performance under different RAM scenarios

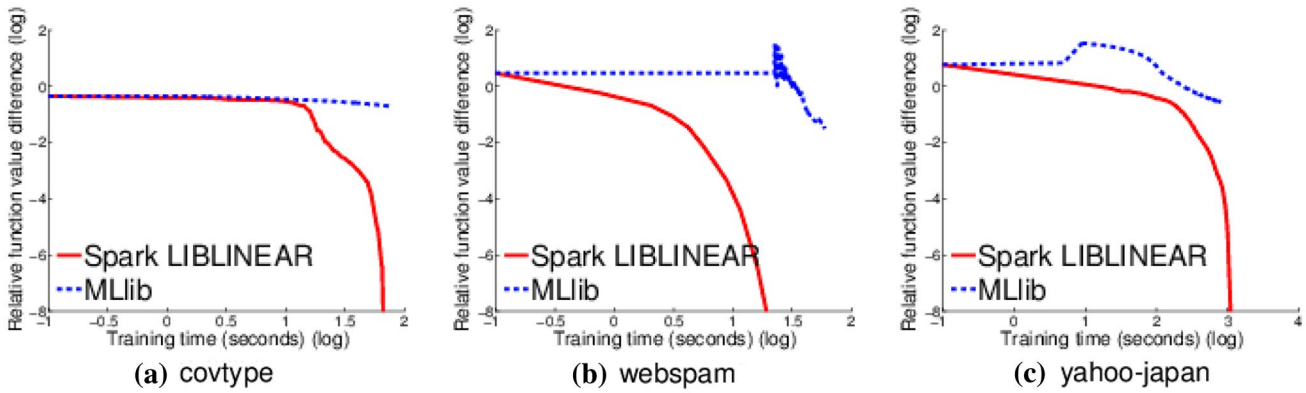

Fig. 7 The performance under different data size scenarios

3.3 Experiment

In this section, we conduct the experimental analysis for the proposed model. In order to test the effectiveness of the proposed solution, multiple modes are tested separately. In the following, two microphones are placed linearly in parallel, and the distance between the signal input port and the sound source is about 8 cm. Audio signals in all the directions are 24 bit floating point number wav files. After being separated by FastICA algorithm, the output is two separated 24 bit integer wav files.

In the Fig. 6, we present the performance under different RAM scenarios.

According to the evaluation index system of the intelligent voice assistant's knowledge service ability, this research selects the corresponding questions for each secondary finger, and builds a question bank for the evaluation of the intelligent voice assistant's knowledge service ability. To ensure the reliability of the test results, the source of the question bank is related books, tests, and thesis. Based on the knowledge map of Internet of things edge equipment, this paper then studies the sharing and collaborative management of Internet of things edge resources, the scheduling and task allocation of resources, the spontaneous self-organization of Internet of things edge computing resources, and constructs an automated and autonomous Internet of things software definition edge computing platform to realize flexible management and intelligent collaboration of Internet of things edge resources and the sharing and dynamic reconstruction of Internet of Internet of things edge services. In the Fig. 7, we present the performance evaluation under different data size scenarios. The experiment is deployed on the Matlab simulation platform. It contains a total of 100 nodes, and the nodes are distributed in an area of 100 m × 100 m. Each node can directly communicate with all nodes in a circle with its own point and a radius of 30 m. The distribution of nodes is random, but the random environment used in all experiments is the same. Set the time period to 20 s.

4 Conclusion

Fusion effect of SVM in Spark architecture for the speech data mining in cluster structure is studied in this manuscript. In terms of the intelligent edge computing framework, the framework of the edge computing task coordination and resource scheduling, power sharing and also collaborative management framework, software definition edge computing framework and equipment collaborative rule engine are proposed. In the research of edge computing intelligent algorithm, the edge-side multi-source data intelligent fusion analysis algorithm, the analysis algorithm based on heterogeneous data and the behavior recognition algorithm in the field of the intelligent speech analytic framework are proposed. Through the experimental analysis, the results have proven the satisfactory performance. In the future research, we will take special attention to the robustness of the model.

References

Alassaf, N., Gutub, A., Parah, S. A., & Al Ghamdi, M. (2019). Enhancing speed of SIMON: A light-weight-cryptographic algorithm for IoT applications. *Multimedia Tools and Applications, 78*(23), 32633–32657.

Chauhan, D. S., Singh, A. K., Kumar, B., & Saini, J. P. (2019). Quantization based multiple medical information watermarking for secure e-health. *Multimedia Tools and Applications, 78*(4), 3911–3923.

Chen, Q., Zhang, G., Yang, X., Li, S., Li, Y., & Wang, H. H. (2018). Single image shadow detection and removal based on feature fusion and multiple dictionary learning. *Multimedia Tools and Applications, 77*(14), 18601–18624.

Cordero, J. A., Nebro, A. J., Barba-González, C., Durillo, J. J., García-Nieto, J., Navas-Delgado, I., et al. (2016). Dynamic multi-objective optimization with jmetal and spark: A case study.

International workshop on machine learning, optimization, and big data (pp. 106–117). Cham: Springer.

Deng, W., Yao, R., Zhao, H., Yang, X., & Li, G. (2019). A novel intelligent diagnosis method using optimal LS-SVM with improved PSO algorithm. *Soft Computing, 23*(7), 2445–2462.

Gupta, A., Thakur, H. K., Shrivastava, R., Kumar, P., & Nag, S. (2017, November). A big data analysis framework using apache spark and deep learning. In 2017 IEEE international conference on data mining workshops (ICDMW) (pp. 9–16). IEEE.

Ibrahim, F., El-Gindy, S. A. E., El-Dolil, S. M., El-Fishawy, A. S., El-Rabaie, E. S. M., Dessouky, M. I., et al. (2019). A statistical framework for EEG channel selection and seizure prediction on mobile. *International Journal of Speech Technology, 22*(1), 191–203.

Kadyan, V., Mantri, A., Aggarwal, R. K., & Singh, A. (2019). A comparative study of deep neural network based Punjabi-ASR system. *International Journal of Speech Technology, 22*(1), 111–119.

Lang, S. M., Bernhardt, T. M., Bakker, J. M., Yoon, B., & Landman, U. (2018). The interaction of ethylene with free gold cluster cations: Infrared photodissociation spectroscopy combined with electronic and vibrational structure calculations. *Journal of Physics: Condensed Matter, 30*(50), 504001.

Maleki, N., Loni, M., Daneshtalab, M., Conti, M., & Fotouhi, H. (2019). SoFA: A spark-oriented fog architecture. In IECON 2019-45th annual conference of the IEEE industrial electronics society (Vol. 1, pp. 2792–2799). IEEE.

Ning, J., Yang, J., Jiang, S., Zhang, L., & Yang, M. H. (2016). Object tracking via dual linear structured SVM and explicit feature map. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4266–4274).

El Ouahabi, S., Atounti, M., & Bellouki, M. (2019). Toward an automatic speech recognition system for amazigh-tarifit language. *International Journal of Speech Technology, 22*(2), 421–432.

Ozcan, T., & Basturk, A. (2019). Transfer learning-based convolutional neural networks with heuristic optimization for hand gesture recognition. *Neural Computing and Applications, 31*(12), 8955–8970.

Ray, R. B., Kumar, M., & Rath, S. K. (2016). Fast computing of microarray data using resilient distributed dataset of apache spark. *Recent advances in information and communication technology 2016* (pp. 171–182). Cham: Springer.

Sangaiah, A. K., Medhane, D. V., Han, T., Hossain, M. S., & Muhammad, G. (2019). Enforcing position-based confidentiality with machine learning paradigm through mobile edge computing in real-time industrial informatics. *IEEE Transactions on Industrial Informatics, 15*(7), 4189–4196.

Singh, T., Di Troia, F., Corrado, V. A., Austin, T. H., & Stamp, M. (2016). Support vector machines and malware detection. *Journal of Computer Virology and Hacking Techniques, 12*(4), 203–212.

Song, T., Pang, S., Hao, S., Rodríguez-Patón, A., & Zheng, P. (2019). A parallel image skeletonizing method using spiking neural P systems with weights. *Neural Processing Letters, 50*(2), 1485–1502.

Souza, M. A., Miyake, H., Borello-Lewin, T., da Rocha, C. A., & Frajuca, C. (2019). α-Cluster structure above double-shell closures and α-decay of 104Te. *Physics Letters B, 793*, 8–12.

Sreeyuktha, H. S., & Reddy, J. G. (2019). Partitioning in apache spark. *Innovations in computer science and engineering* (pp. 493–498). Singapore: Springer.

Suthakar, U., Magnoni, L., Smith, D. R., & Khan, A. (2016). Optimised lambda architecture for monitoring WLCG using spark and spark streaming. In 2016 IEEE nuclear science symposium, medical imaging conference and room-temperature semiconductor detector workshop (NSS/MIC/RTSD) (pp. 1–2). IEEE.

Śmieja, M., & Wiercioch, M. (2017). Constrained clustering with a complex cluster structure. *Advances in Data Analysis and Classification, 11*(3), 493–518.

Talan, P. P., Sharma, K. U., Nawade, P. P., & Talan, K. P. (2019). An overview of hadoop MapReduce, spark, and scalable graph processing architecture. *Recent developments in machine learning and data analytics* (pp. 35–42). Singapore: Springer.

Thakur, S., Singh, A. K., Ghrera, S. P., & Elhoseny, M. (2019). Multilayer security of medical data through watermarking and chaotic encryption for tele-health applications. *Multimedia Tools and Applications, 78*(3), 3457–3470.

Vapnik, V., & Izmailov, R. (2017). Knowledge transfer in SVM and neural networks. *Annals of Mathematics and Artificial Intelligence, 81*(1–2), 3–19.

Wang, W., Lilyestrom, W. G., Hu, Z. Y., & Scherer, T. M. (2018). Cluster size and quinary structure determine the rheological effects of antibody self-association at high concentrations. *The Journal of Physical Chemistry B, 122*(7), 2138–2154.

Wottschel, V., Chard, D. T., Enzinger, C., Filippi, M., Frederiksen, J. L., Gasperini, C., et al. (2019). SVM recursive feature elimination analyses of structural brain MRI predicts near-term relapses in patients with clinically isolated syndromes suggestive of multiple sclerosis. *NeuroImage: Clinical, 24*, 102011.

Wu, X., Zuo, W., Lin, L., Jia, W., & Zhang, D. (2018). F-SVM: Combination of feature transformation and SVM learning via convex relaxation. *IEEE Transactions on Neural Networks and Learning Systems, 29*(11), 5185–5199.

Xiong, X., Tang, R., & Yang, X. (2019). Finite-time synchronization of memristive neural networks with proportional delay. *Neural Processing Letters, 50*(2), 1139–1152.

Yu, Z., Zhu, X., Wong, H. S., You, J., Zhang, J., & Han, G. (2016). Distribution-based cluster structure selection. *IEEE Transactions on Cybernetics, 47*(11), 3554–3567.

Zhang, S., Wang, H., & Huang, W. (2017). Two-stage plant species recognition by local mean clustering and Weighted sparse representation classification. *Cluster Computing, 20*(2), 1517–1525.

Zhang, S., Wang, H., Huang, W., & You, Z. (2018). Plant diseased leaf segmentation and recognition by fusion of superpixel, K-means and PHOG. *Optik, 157*, 866–872.