TEST Engineering & Management

January-February 2020
ISSN: 0193-4120 Page No. 8363 - 8370

Diseases Classification with Genetic Algorithm for Support Vector Machine using Hadoop

Prof. Shrikant P. Akarte, Assistant Professor, Dept. of CSE, PRMIT&R, Badnera,India.
Dr. G. R. Bamnote, Professor & Head, Dept. of CSE, PRMIT&R, Badnera,India

Article Info
Volume 82
Page Number: 8363 - 8370
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 07 February 2020*

Abstract:
This Paper proposes classification method based genetic algorithm (GA), for many real imbalanced data sets, which has a very small number of different objects and a large number of certain type objects. Support vector machine (SVM) which is a normal classification method, dose not work for skewed statistics sets. Mainly this work is focused on classification of medical disease by the combining hereditary algorithm and uphold vector machines (SVM) which is a feature selection technique higher performance, SVM is best as compared to conventional learning steps within applications. SVM is relatively a novel classification technique. To increase the overall performance of SVM, we use combination of GA and SVM. The proposed method has better classification accuracy related to further admired classification algorithms for several skewed data sets.

Keywords:*Classification, Genetic Algorithm, Hadoop, Support Vector Machines, GA-SVM.*

I. INTRODUCTION

In medical mining technology due to recent advances, the medical data classification has become more challenging problem. For the classification of tremendous amount of medical data which takes lots of time and also takes excessive computational effort, for that many applications may not be appropriate. During this work, to optimize the support vector machine parameters, we have a tendency to combine the sensible points of genetic algorithmic program (GA) and (SVM).

Classification is most important technique used for data mining in biomedical research. Examples of training are required to foretell a target class of an unknown example in the classification task. Nevertheless, training data sometimes have imbalanced class distribution. Insufficiency of total quantity of exemplars of a few classes for training a classifier is great measure of this problem. [1].

The combination of GA and SVM is able to select "optimal feature set" and also able to evaluate "optimal parameters" for SVM classifier. For looking out in massive search areas GA is one in all the foremost powerful tools and it imposes few mathematical constraints within the form of the operate to optimize [3]. Merging of Genetic Algorithm and Support Vector Machine is incredibly unique technique, such that many researchers are working on the combination to improve the accuracy of classification in Support Vector Machine [2, 4].

In this paper, we join GA with SVM to take care of the grouping issue of slanted medicinal informational collections. The distinction with the other imbalanced information order is the information points generated by GA and bolster vectors are optimal. So the data joined is reasonable and the arrangement precision is improved. We first use SVM to prepare the entire information and discover the help vectors (SV). At that point we use GA to produce new information close SV and the choice limit until the presentation foundation is improved. A few benchmark imbalanced medicinal informational indexes are utilized to contrast our calculation and the others. The outcomes show that the arrangement exactness is improved uncommon.

January-February 2020
ISSN: 0193-4120 Page No. 8363 - 8370

II. LITERATURE REVIEW

A. Im-balanced Data

Many researchers have studied the problem of imbalanced data classification to improve output of categorization model. Previous researches mainly worked related to binary classification. Steps in algorithm such as feature selection and the factors for algorithm such as sampling-based approach & cost sensitive learning can be extended to multiclass problem but at the same time it is difficult to apply for algorithm specific techniques. [5]. There are many research works that try to improve traditional techniques or develop new algorithms to solve the class imbalance problem but their research as stated limited to binary classifier. Only a few researches have been done for multiclass imbalance problem that is much more common and complex in the real world application.

We will study the multiclass imbalanced data problem, and developed new classification algorithms that can effectively handle the imbalance problem in many biomedical domains. Authors implemented the algorithm of multi class kernel-dependent vector machines. Wasikowski M, Chen X W.worked on the smallsample class imbalance problem using feature selection. Similarly, Chen X, Gerlach B, Casasent D introduced support vectors for imbalanced data classification [6].

B. Data Mining

Past researchers have defined “Data Mining” as the disclosure of "models" for information. This "model" can be a few things. The endeavor to blackmail data which was not acknowledged through the information is designated "information mining" or "information digging". For instance: Suppose our information is a lot of numbers. This information is much effectively comprehended than information that would be information mined, however it will serveas an example. Gaussian dispersion gives information and this information utilize an equation to figure the most comparable parameters of this Gaussian. The information become the model of the information, which originate from mean and standard deviation of this Gaussian dissemination totally describe the dispersion. Presently days, most PC researchers said that information mining as an algorithmic issue. In such case, the model of the information is just the response to a perplexing inquiry about it. There are a wide range of ways to deal with displaying information [8].

C. Hadoop Overview

Hadoop is a distributed computing framework released by Apache Foundation, it is Google's open source implementation of the cloud computing model, and it can be efficient, reliable, scalable way to process data. Its core ideais to build on a large number of cheap and efficient cluster hardware devices, in the form of software processing to provide storage and computing environment for the huge amounts of data, and provide a unified standard interface, is a highly scalable distributed computing systems. Hadoop have technique of Map and Reduce two programming functionalities to handle mass of data. In past due to hadoop, there are several cloud computing based simulation system is developed such a calculation based on the concept of cloud modeling and simulation platform of COSIM-CSP system, a new mode of the networked manufacturing, private cloudframework for visual simulation, and the military trainingsystem [7].

D. Classification

Classification is defined as the assigning of object to predefined categories. It contains much diverse application, which has pervasive problem. For example, detecting spam email messages depend upon content and message header, classifying galaxies depend upon their shapes and categorizing cells as malignant or benign depend upon the results of MRI scans. Collection of records is an input data for a classification task. Each record, also called as example or instance, which then specified via a tuple (x, y), group label contains x and y, where x is known as the attribute set and y is known as a special attribute. For the purpose of classification, the attributes set in a selected dataset can be of two types: discrete or continuous. Class labels must be

January-February 2020
ISSN: 0193-4120 Page No. 8363 - 8370

continuous in nature. Regression, a predictive modeling task in which y is a continuous attributes is the key characteristic that distinguishes classification [9].

E. Multiclass Classification

III. In multiclass grouping, given a lot of marked models with names chose from a limited set, an inductive method assembles a capacity that (ideally) can delineate examples to their suitable classes.

IV. PROJECTED SYSTEM

Block diagram of proposed approach is shown in figure 3.1. a) Initially, start all the Hadoop clusters to start Namenode, Secondary Name Node, Data Node, Task Tracker and Job Tracker.

b) Take input csv file which contains different diseases data then copy that csv file into HDFS

c) Apply GA-SVM to csv file. In GA-SVM, procedure standards are recognized via structure and output part file is created

d) Convert that output part file into csv file for Neo4j, then finally load csv file into Neo4j for analysis.

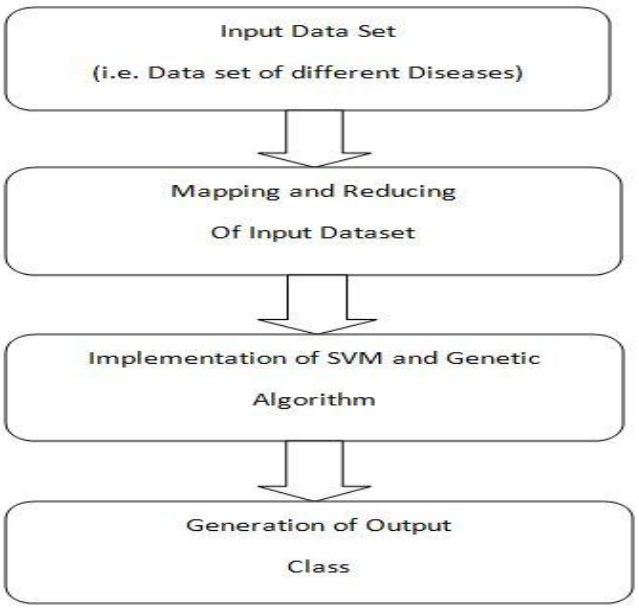


Figure 3.1. Proposed System Architecture

V. METHOD IMPLEMENTED

We depicted the projected Genetic-SVM structure for the characteristic choice. The aim of this system is to optimize the SVM classifier. For this, we have to pick the subset of options mechanically.

A. Genetic Setup

The first step in GAs is to stipulate the key writing allowing describing any potential answer as a numerical vector, we are supposed to use vector of (0 and 1) with span of twenty two (types of options) that 0 & 1 is for the eliminated and elite options severally. At initial, willy-nilly we got a bent to come up with fifty chromosomes as a population.We have a tendency to use circle choice for the cross- over with to boot we've got a bent to use exchange. This machinist merely allow to change location of two samples arbitrarily. Then likelihood constraint of alteration be equivalent to zero.1. The selection of the fitness operate is extremely necessary as a results of basis that the Genetic calculates the decency of every contender declare coming up with SVM structure.

B. SVM categorization through genetic algorithm

SVM classification process is as follows:

Step 1. Initially generates size of hundred populations randomly.

Step 2.Training SVM Classifier: With the variable value of parameters and selected feature subset, SVM classifier is trained.

Step 3. For computing each chromosome (subset of features) fitness, train (n(n-1))/2 SVM Classifier.

Step 4. Renew new individuals from old ones and based on fitness value, select individuals from population.

Step 5. The variety of iteration is not reached to extent isn't however reached, then we move further to next generation process. The extinction criteria are severe bodily harm variety reached or the strength operate price will not get better throughout the previous fifteen generations come back to step a pair of.

Step 6. Select the best option as optimum set characteristic.

Step 7. Apply the optimal option to dataset.

January-February 2020
ISSN: 0193-4120 Page No. 8363 - 8370

Step 8.Genetic Operation. The replica operators chosen between two hundredth of the best body.

VI. Experimental Analysis

A. Requirement Analysis

For the implementation of this system, we used Ecillips IDE with Hadoop. In computer programming, Eclipse is nothing but an integrated development environment (IDE), used for categorizing the surroundings, it has extended plug-in structure and a bottom workstation. Eclipse may be worn to widen applications, this applications written mostly in java. Eclipse is a software development kit (SDK), it contains Java development tools, for java developers. In Eclipse user can contribute and inscribe their individual connect modules and also contains plug-ins written for the Eclipse Platform, such as development toolkits for other programming languages. This plug-ins provides all the functionality within and on top of the runtime system. Also we used Hadoop, Apache Hadoop is a framework which works across clusters of commodity computers using a simple programming model for the distributed processing of large data sets. Hadoop is designed to broaden from single servers to thousands of machine servers and each of them is endow with computation as well as storage.

B. Hardware and Software Requirements

- Software Environment

The system will run under Eclipse Framework that is to be installed on the system.
Operating Platform : Ubantu 14.04 LTS
Front End : Eclipse Luna-SDK
File System : HDFS

- Hardware Environment

CPU : Dual Core or beyond
Random access memory: 4 GB RAM
Hard Disk: 40 GB
LAN : Enable

VII. Result

First, Dataset is selected for the purpose of classification. Next step is preprocessing i.e. removing unwanted data.

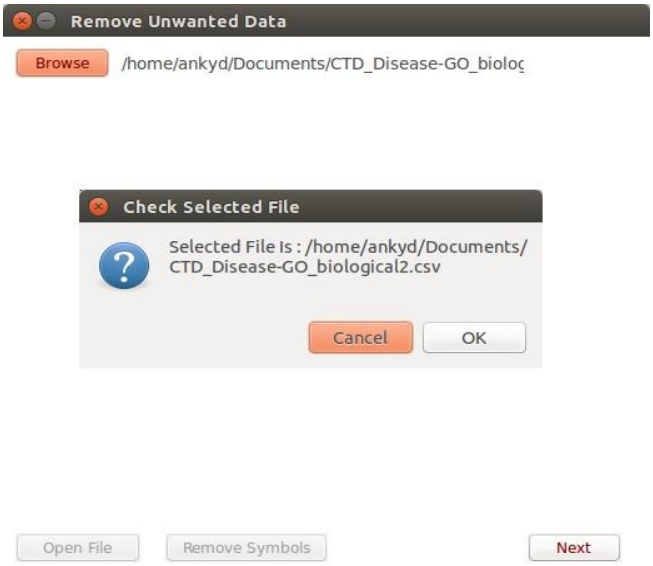


Figure 6.1. collection of statistics

Removing unwanted data i.e. preprocessing is important for the accuracy of classifier. In this step, unnecessary symbols, whitespace, stop words etc. in dataset is removed.

Figure 6.2. Preprocessing of Data

January-February 2020
ISSN: 0193-4120 Page No. 8363 - 8370

After removal of unnecessary data which consists of symbols, punctuations, stop words, whitespace, converting upper case letters into small letters is done then this new dataset is stored and selected by system for next step.

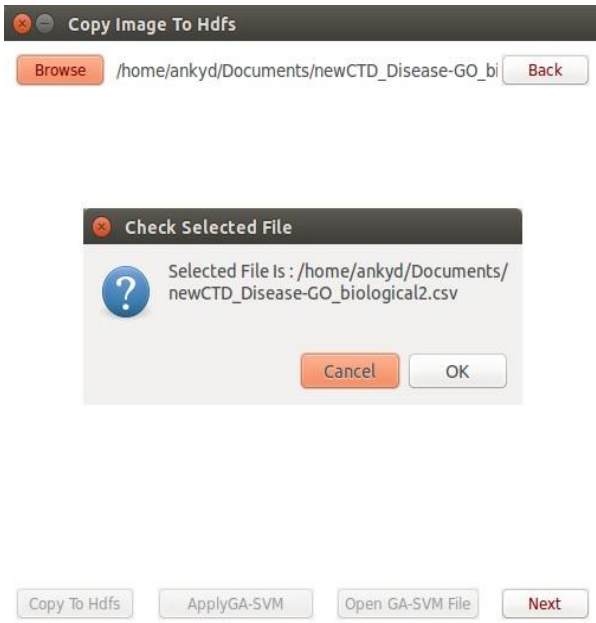


Figure 6.3. Selection of preprocessed data set

As hadoop is scalabe way to processed data, file need to copied to HDFS(Hadoop Distributed File System)

Figure 6.4. Copied file to HDFS

Next step is to apply Genetic Algorithm based support vector machine algorithm for the purpose of classification.

We planned and put into practiced genetic algorithm (GA) to optimize kernel parameters for Support vector machine. Another advantage of implementing Genetic algorithm is feature subset selection for SVM classification and applied it to the classification of disease.

Figure 6.5. Apply GA-SVM

Map Reduce is technique to handle big amount of data. Its consist functions: Mapper and Reducer. Background process of these functionalities is shown in following figure.

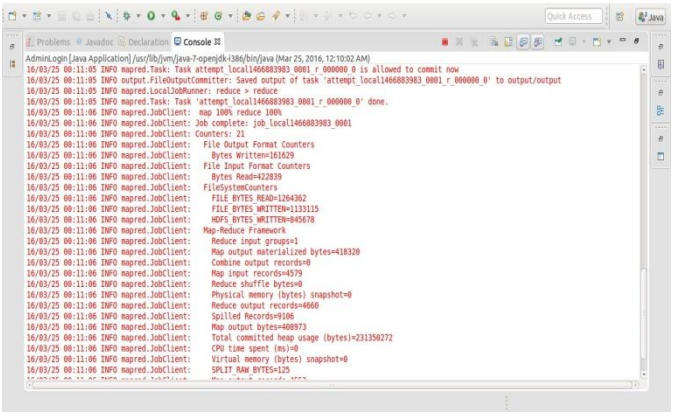
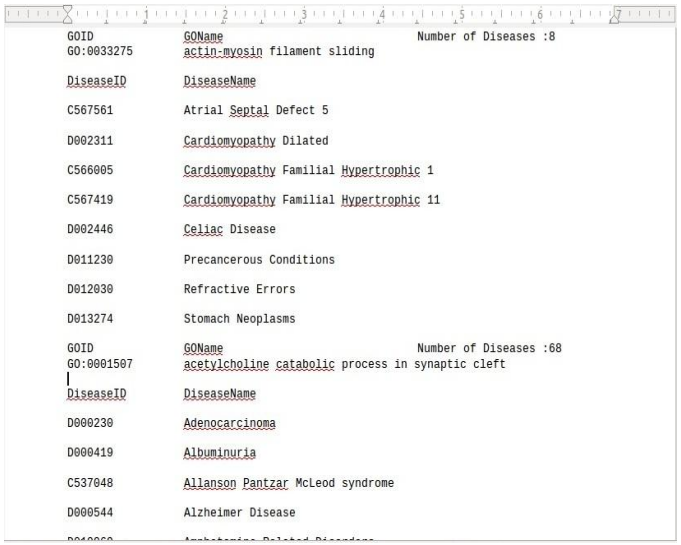


Figure. 6.6. Background Process of Map Reduction &GA-SVM

January-February 2020
ISSN: 0193-4120 Page No. 8363 - 8370

In GA-SVM, method standards are recognized through the system and output part file is created. Output Part file contains classification of Diseases



```
GOID            GOName                          Number of Diseases :8
GO:0033275      actin-myosin filament sliding

DiseaseID       DiseaseName

C567561         Atrial Septal Defect 5

D002311         Cardiomyopathy Dilated

C566005         Cardiomyopathy Familial Hypertrophic 1

C567419         Cardiomyopathy Familial Hypertrophic 11

D002446         Celiac Disease

D011230         Precancerous Conditions

D012030         Refractive Errors

D013274         Stomach Neoplasms

GOID            GOName                          Number of Diseases :68
GO:0001507      acetylcholine catabolic process in synaptic cleft

DiseaseID       DiseaseName

D000230         Adenocarcinoma

D000419         Albuminuria

C537048         Allanson Pantzar McLeod syndrome

D000544         Alzheimer Disease
```

Figure 6.7. Output Part File

For further analysis step, this output part file is needed to convert into comma separated value format.

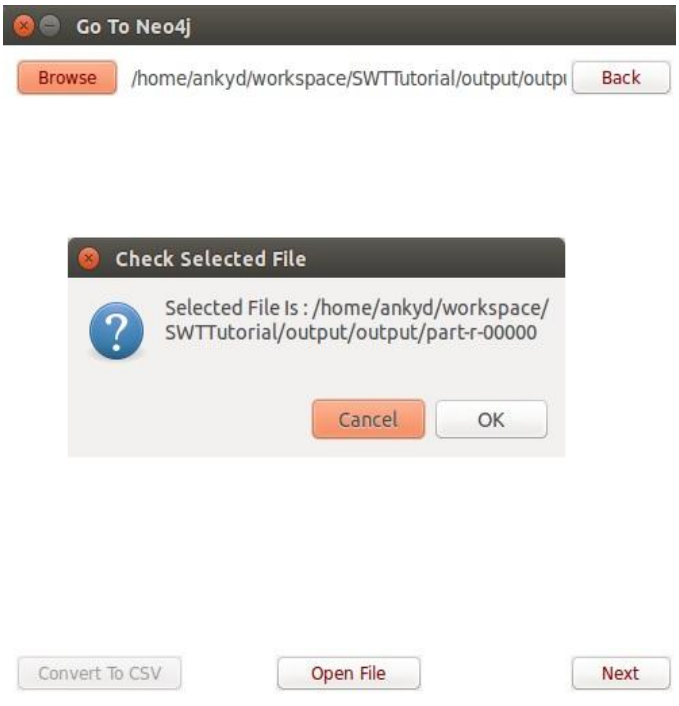


Figure 6.8. Get Part file to convert into csv for Neo4j

Figure 6.9. CSV file created for Neo4j

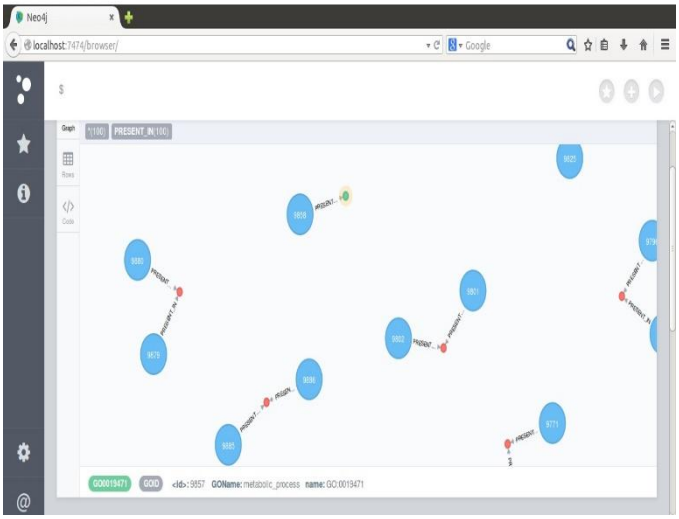


Fig. 6.10.Graph structure creation for classification of Diseases

VIII. COMPARATIVE ANALYSIS

with consideration to the projected algorithm toward range of data sets results obtained. intended for testing, two standard datasets are taken from the Comparative Toxico genomics Database as shown in Table 7.1. SVM, SVM with Genetic classification techniques are used to validate the prediction results.

Table 7.1: Datasets used for comparison

Datasets
CTD_Disease-GO_biological_process_associations
CTD_diseases_pathways

January-February 2020
ISSN: 0193-4120 Page No. 8363 - 8370

The performance of a proposed classification is found out using two factors: Computation time and Error rate. The categorization correctness is understood by means of compassion and uniqueness. The calculation instance is written in favor of two classifier is considered in account.

The assessment characteristics is the uniqueness, compassion, and taken as a whole accuracy of five things of data sets are offered in Table 7.2 and Table 7.3. Proposed and existing SVM correctness and fault rates are shown in figure 7.1 and figure 7.2.

Three traditional assessment rules of accuracy, remember and F-score are worn to assess the effectiveness of the projected technique. To perform classification positive and negative classes are used with the three metrics. Positive predictions that are correct is equals to precision and positive samples that are correctly predicted positive is equals to recall. That is:

Accuracy = T+ve ÷ (T+ve + F+ve)
Recall = T+ve ÷ (T+ve + F-ve)
F − score = (2 ∗ Accuracy ∗$Recall$) ÷ (Accuracy + Recall)

- True +ve =correctly predicted number of +ve samples.
- False -ve (FN) =wrongly predicted number of +ve samples.
- False +ve (FP) =wrongly predicted number of -ve samples as positive.
- True -ve (TN) =correctly predicted number of -ve samples.

Table 7.2: presentation of various data sets SVM

Datasets	TP Rate	FP Rate	Precision	Recall	F-Measure
CTD_Disease-GO_biological_process_associations	0.954	0.09	0.954	0.954	0.954
CTD_diseases_pathways	0.774	0.34	0.766	0.774	0.762

Table 7.3: presentation of various data sets SVM with Genetic

Datasets	**True+ve Rate**	**False +ve Rate**	**Precision**	**Recall**	**F-Measure**
CTD_Disease-GO_biological_process_associations	0.967	0.07	0.967	0.967	0.967
CTD_diseases_pathways	0.773	0.335	0.769	0.773	0.77

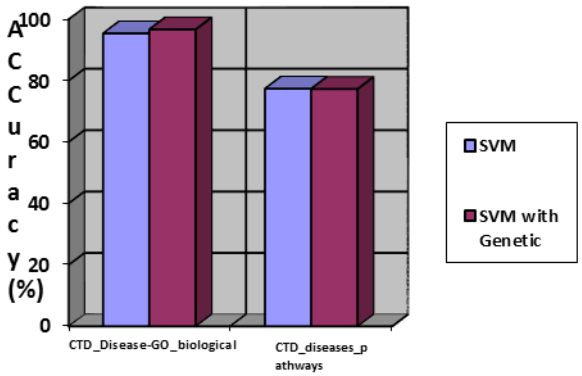


Figure 7.1.Compares accuracy by the genetic-SVM way with plain SVM classifier

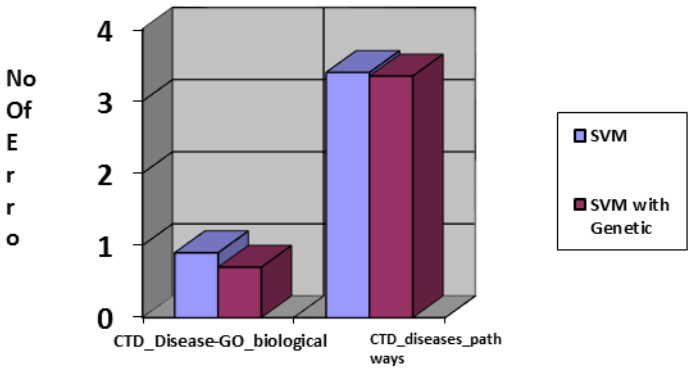


Figure 7.2. Error rate

IX. CONCLUSION

This article projected a Genetic algorithm depend on optimization algorithm, that can minimize the kernel parameter standards for SVM, and get the minimal subset of qualities. Further SVM with genetic algorithm is planned and implemented to eliminate unrelated features and competently find out most excellent parameter principles. Main objective of this article is to design sustain Vector

January-February 2020
ISSN: 0193-4120 Page No. 8363 - 8370

Machine and Genetic Algorithm were calculated en route for finding the categorization accurateness with runtime intended for different kernel functions such as Polynomial and Radical essential function are worn. Optimal characteristic range algorithm is necessary for correctness of algorithm by means of high opinion to remedial datasets which we considered. The outcome shows the categorization accurateness of GA-SVM is the higher than traditional SVM algorithm.

REFERENCES

1. Piyaphol Phoungphol__, Yanqing Zhang, Yichuan Zhao. Robust Multiclass Classification for Learning from Imbalanced Biomedical Data.
2. Xue-wen Chen, "Gene selection for cancer classification using bootstrapped genetic algorithms and support vector machines", In 2nd IEEE Computer Society BioinformaticsConference (CSB 2003), 11-14 August 2003, Stanford, CA, USA, pp. 504-505.
3. Melanie Mitchell, An Introduction to Genetic Algorithms, MIT Press, 1996.
4. Frohlich, H, et al, "Feature selection for support vector machines by means of genetic algorithm", In 15th IEEEInternational Conf. on Tools with Artificial Intelligence, 2003, pp. 142-148.
5. Chawla N V, Japkowicz N. Editorial: Special issue on learning from imbalanced datasets. SIGKDD Explorations, 2004, 6: 1-6.
6. Y. Liu et al., "Combining integrated sampling with SVM ensembles for learning from imbalanced datasets," Information Processing & Management, vol. 47, no. 4, pp. 617-631, Jul, 2011.
7. hadoop.apache.org, Apache Foundation.
8. S¸eyda Ertekin1, Jian Huang, L´eon Bottou, C. Lee Giles "Active Learning in Imbalanced Data Classification."