National College of Ireland

Higher Dimensional Feature Reduction using Hybrid Particle Swarm Optimization

MSc Research Project
Data Analytics

Jayaprakash Subramanaim
x15009866

School of Computing
National College of Ireland

Supervisor: Dr. Jason Roche

National College of Ireland
Project Submission Sheet – 2015/2016
School of Computing

Student Name:	Jayaprakash Subramanaim
Student ID:	x15009866
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Dr. Jason Roche
Submission Due Date:	21/12/2016
Project Title:	Higher Dimensional Feature Reduction using Hybrid Particle Swarm Optimization
Word Count:	4086

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	21st December 2016

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Higher Dimensional Feature Reduction using Hybrid Particle Swarm Optimization

Jayaprakash Subramanaim
x15009866
MSc Research Project in Data Analytics

21st December 2016

Abstract

Digital informations are tremendously increasing at every second, in the recent decade there has been a surplus amount of data storing day by day. With the advancement in efficient machine learning algorithms, analyzing of complex data becomes much easier. Categorizing these textual data recently became more important for the organizations dealing with huge data. Handling higher dimensionality features becomes more complicated on process huge data, forcing to select certain Features (FS) which are higher values for the process. In this project an efficient feature selection algorithm which is based on Hybrid Particle Swarm Optimization (PSO) is used improve the efficiency of Feature Selection process. Here Support Vector Machine is used as the classifier in combination with PSO to give a best features.

1 Introduction

The 21st century is called the era of Internet, where informations are shared around the globe at every second. Most of the informations available are in the form of unstructured data, it is also said more than 80% of the informations are available are unstructured data VS and P (2013). These data are generated mostly from social media such as Facebook, Twitter, bloggers, News article, review sites, etc which are more valuable for the industries depending on it. Inorder to process those data and convert them into valuable information researches has developed many algorithm. The research question is based on implementation of an algorithm which can solve a higher dimensional feature reduction in terms of classification various data.

Among many algorithms Particle Swarm Optimization provides a best results StuartReid (n.d.) for search base problems and optimization problems. Since the research proposal is based on higher dimensional feature reduction, a Hybrid Particle Swarm Optimization is used to reduce the higher dimensional features in the multi- dimensional space vector and categorize the data accordingly to its domain using Support Vector Machine. Furthermore, this research will analysis the various optimization techniques Zhang et al. (2014) involved in Particle Swarm Optimization algorithm. The composition of this research paper are as follows: Section 2 will discuss the similar and supporting works carried out on the research area. Section 3 will discuss techniques used in this project and

the type of data used for the analysis. Section 4 will be explain the implementation process of the techniques. Section 5 will be on evaluating the results produced in the section 4. Finally section 6 will be on findings where futuristic research can be carried on.

2 Related Work

Text feature selection is one of the major preprocessing stage which is carried out in text mining and text categorization. Vector space model (VSM) Salton and McGill (1983) is basic model used for Text categorization which involves usage of different categories as classes and words in the document as features. There are many methods which can be used to generate the value for features such as Term Frequency (TF), odd-ratio, CHI-Square ratio Prusa et al. (2015), document frequency (DF), Mutual Information (MF), etc. Improvement of weight Dong and Hownet (n.d.) are can be done by using lexicon and semantic models. As there are many parameters to take care of during the conversion many researchers suggest to opt for a mixed methodology for filtration criteria such as Term Frequency Inverse document frequency (TF-IDF) Salton and Buckley (1988) and Inverse Category Frequency (ICF) How and Narayanan (2004).

Studies has suggested usage of wrapper model in analysis of higher dimensional feature space does not provide an promising results as expected in terms of performance Yang and Pedersen (1997). So latent semantic index (LSI) can be used for transform by implementing text in feature space vector, which helps for a higher dimensional feature problems. With the help of this approach filtration of required feature which contribute high efficiently to the process can be taken out from higher dimensional feature space. By implementing Genetic Algorithm (GA), searching even a fine grain of features provides a better results l.G. Hao and l.O. Wang (2013).

An evolution of LSI and GA lead to the development Particle Swarm Optimization (PSO), which is a population based optimization technique developed by Kennedy and Eberhart in 1995 Kennedy and Eberhart (1995). Particle Swarm works on the population based search related problems, it initializes with population of particles in N Dimensional space vectors. This methodology is used in detection of spam messages from other normal messages Lai and Wu (2007).

The main drawback of Genetic Algorithm are mutation and crossover Hu (n.d.), so overcome this issue Particle Swarm Optimization model is used. Working of PSO is much similar to bird flocking mechanism, where they move in a uniform fashion in search of food helping them to cover more area in single iteration. As the number of iteration increases the area in search of food decreases. So this mechanism was first developed as Ant colony optimization Al-Ani (7) where the ant move in uniform faction to find the food. Particle Swarm optimization is introduced inorder to overcome the issue produced by Ant Colony optimization algorithm which is on collision avoidance. Later PSO became a common methodology in the field of optimization and Feature Selection. Grid based search method in early stage was replaced with PSO which including an classifier and Feature selection algorithm at the same time Lin et al. (2008). Even the famous Hill-climbing method was replaced with PSO in selection of features based on rough sets Wang et al. (2007). Later on there where different types of PSO where introduced in resolving various different problems according to the needs such as Binary or Basic Particle Swarm Optimization (BPSO), Simple Particle Swarm Optimization (SPSO), Extreme Distributed and Simple

Particle Swarm Optimization (TSPSO) and Dynamic Extremely Distributed and Simple Particle Swarm Optimization (DTSPSO). Here the gaps of each algorithm are updated with the later version. As The velocity of particle movement from BPSO is updated in SPSO, where as in TSPSO threshold of disturbance is taken into considerings in updating the particle position.

There were many supervised and unsupervised methods introduced along with PSO in classification problems. Clustering base algorithm with K-mean was combined with PSO as a hybrid PSO in document clustering provided a good results Cui et al. (2005). Later when considering a classification based problems SVM provides and positive results with PSO other than decision tree, Artificial neural networks [Feature Selection for Chinese Text Categorization Based on Improved Particle Swarm Optimization]. In order to fine grain of the SVM classification process Transductive Support Vector Machine (TSVM) and simulated annealing (SA) was introduced to improve the standards Sun and Sun (2005). For this research topic a classification model known as LIBSVM Chang and Lin (2011)is used to compute 10-fold crosvalidation accuracy, the reason for using LIBSVM is because it has high computational performance which can solve C-SVM classification, epsilon-SVM regression, nu-SVM classification nu-SVM regressions also for multi class classification problems.

3 Methodology

In this research project, the study is focused on the selection of data type and converting those data into useful information by prepossessing and cleaning and finally convert into machine understandable format (i.e. numerical values) be sending it to prediction stage. Once the Data are cleansed up, a sample of data are used for testing of the PSO-SVM algorithm and the accuracy is considered. This sample of data has two classes and 13 top most feature sets. So the calibrations and other parameters are adjusted according to the desired accuracy. The original dataset consist of 5 different classes and 600 feature sets. The technology used in this research process are Python IDLE from Anaconda. Python is one of the most powerful programing languages when it comes to analytics. On the account of considering the project time line factor, machine learning technologies are taken into consideration.

3.1 Data Collection

Since the project is all about Categorization of documents, the datasets are collected from BBC. Yes in this project News feed dataset documents are collected from [http://mlg.ucd.ie/datasets/bbc.html]. This data consists of numerous text documents on each topic providing a training and test datasets. There are basically 5 categories of data classifications/categories Business, Sports, entertainment, politics and technology. These are raw simple news feeds data which appeared at BBC articles on the year 2004 to 205. Summary of the datasets are described in the below table. The reason for choosing these dataset is because of its complexity in natural language processing and diverse data which helps in the analysis of this project.

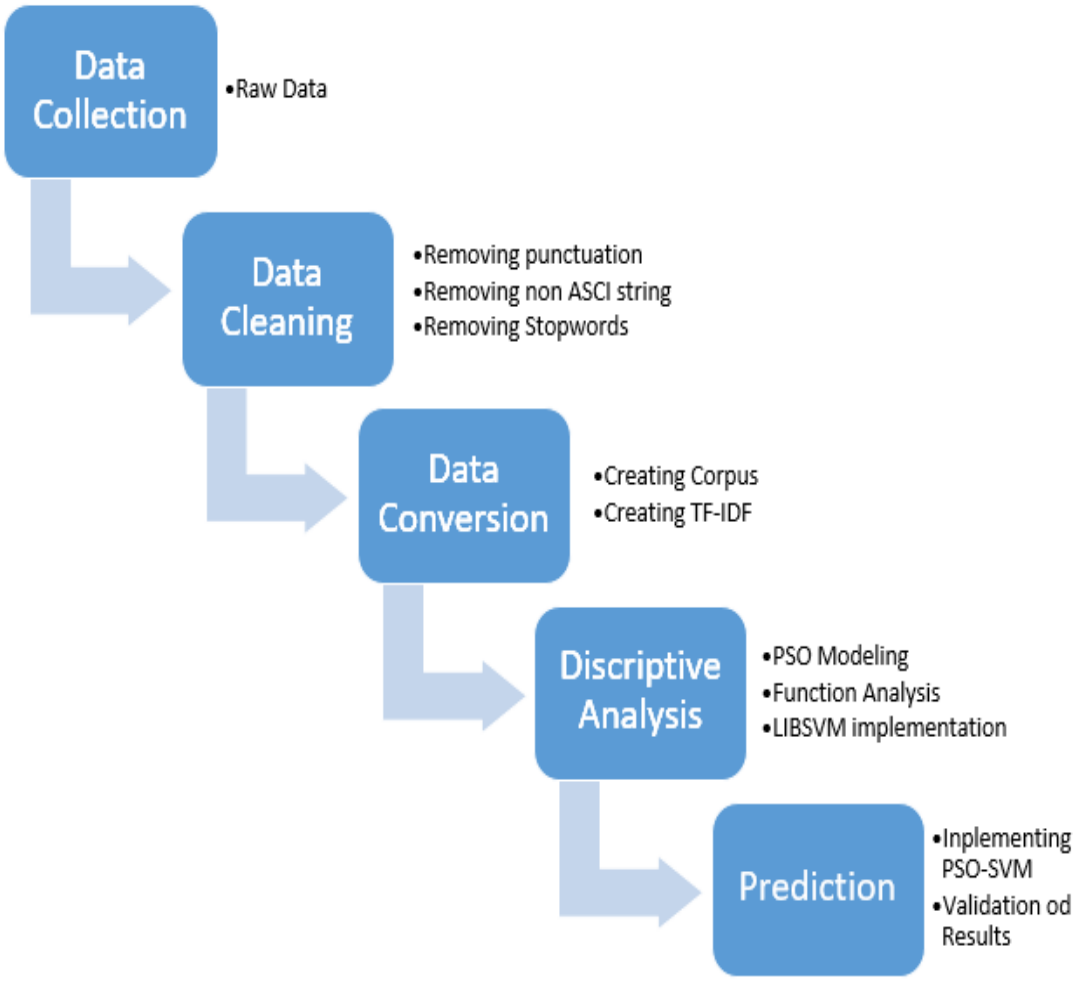


Figure 1: Process Workflow

Topic's	Number of Document
Business	150
Entertainment	386
Politics	417
Sports	511
Technology	401

3.2 Data Cleaning

Data Cleaning is one of the most important stages in the entire process, more to be called as key changing factors. If there is any mislead in this cleaning of data, it will reflect back till the end of the process. So the data Cleaning comprises of basically three stages Punctuation removal, Non ASCI characters removal, Stop words removal. This all process is carried out with a package called Natural Language Tool Kit (NLTK) Bird et al. (2009) in python. Natural language processing is specially used for analysis of text related data, it has preinstalled numerous algorithms especially for text Data Analysis and contains over 50 different corpora and lexical resource. Since this project is not about deep analysis on Text mining, here we are just going to use NLTK for stemming, tokenizing, stopword removal. Basically stemming is a process of reducing inflectional forms of words and derivationally related words. Say example

Having – have
Ponies – pony
Cats —- cat

So first stage in data cleaning is filtering non ASCI characters such as ©, ®, , , , 163, %d, .etc while the document is imported, this help us to do filtrations on related text. Later on goes tokenization, breaking up of words, followed by stopwords removal. This process finally leave the words which are completely cleaned, this can be later used for Term Frequency and Inverse Document Frequency.

3.3 Data Conversion

This is the intermediate stage that is at this stage human understandable language is converted into machine understandable language, ie test is converted into numerical using an algorithm called TF-IDF. TD- IDF B (n.d.)Jain et al. (1999) consists of two stages, one is calculation of Term frequency and Inverse document frequency. From the previous stage, the tokenized word is taken out and the frequency of each word is calculated and stored out in the memory. The most frequent words may sometimes be less useful metric, say example understanding is an unique word which is also not a part of stopwords and it may occur frequently on all documents. The main aim of this stage is to calculate unique words which represents a topic. Hence inorder to measure that, we need to measure how frequently the word occurs in all the documents which is Inverse document frequency. Hence the product of Term frequency and Inverse document frequency gives the weight of the particular word in the entire document. Later the entire data is uploaded to form a corpus to use it for later usage.

$$W_{di} = TF_{di} * IDF_{di} = TF_{di} * log_2(n/DF_{di}) \quad (1)$$

Where

Wdi is the weight of the particular item i across the document d.
TFdi is the number of times the term i occurred in the document d.
DFdi is the term frequency across all documents.
N is the collection of total number of documents.

To make this process simple from NLTK packages TF-IDF algorithm is used and the weight of every item is calculated and saved separately.

3.4 Machine Learning Model Selection

This part of the analysis is the heart of the project, where the data collected from previous stages are combined together in a meaning full way so that it can be directed directly to the machine learning process. This proposal is on usage of Support Vector machine with the combination of Particle Swarm Optimization on the categorization of documents. Document categorization is one of the hottest top on current research fields as it can give a tremendous amount of information. The reason for choosing Particle Swarm Optimization is because of its computing capability when the scale of documents is larger. The optimization capability is high in a very big search space environment, so that in single iteration it can cover up many solutions and obtain a targeted solution in few iterations. Performance of the PSO can be increased by optimizing PSO parameters EVERS (2009). The main advantage of using PSO over other optimization problem is that it does not use gradient of the problem being optimized and need not be differentiable. Say example the most classic optimization method is gradient descendant which need the data to be differentiable. The working mechanism of PSO depends on the population of the particles. These particles move around the search space depending upon the dimension of the solution. Where each particle is treated as a point in N- Dimensional space. The basic working algorithm of the PSO are as follows. Before going deep on the concepts, the simplest form of PSO mechanism can be explained with the bird flocking. As everyone would have seen the movement of birds in a uniform regular fashion in search of food. Initially they dont know where the food is, so they move randomly searching for food in an area. So at each iteration birds move close to the food, though they dont know where exactly the food is. Inorder to find the food, the effective one is to follow which is much closer to the foods. Here PSO uses this scenario to solve the optimization problems, birds are represented as particle and food is represent the solution. At every iteration birds move towards the food by updating two values one is pbest solution and another one is gbest solution. The best values of the fitness from the particle during iteration are stored in the pbest and the best values of global solution are stored in gbest. When the particle is populated to its topological neighbor, the obtained best solution is called local best Hu (n.d.). To find the velocity of the particle in which it has to move to obtain the local position and the position of the particle is shown in equ (2).

$$V_{ij}[] = V_{ij}[] + C1 * rand() * (pbest_{ij}[] - position_{ij}[]) + C2 * rand() + (gbest_{ij}[] - position_{ij}[]) \quad (2)$$

$$Position_{ij}[] = position_{ij}[] + V_{ij}[] \tag{3}$$

where Vij[] is the velocity of the particle obtained
Positionij[] is the current location of the particle
Rand() is the randomly generated number which lies between 0 and 1
C1 and C2 are the constants usually it will be of value 2

Here The standard pseudo code of particle swarm optimization is as follows.

```
Begin:
    Begin populating
    While (stopping condition != True)
        While (iteration <= generation)
            For every particle is the population
                If fitness of particle >Pbest value
                    Set current fitness value to pbest
                For each particle A around fitness of particle
                    If fitness of A is greater than fitness of gbest
                        Set A to gbest
                END
                For each dimension j
                    Calculate equation (1)
                    Calculate equation (2)
                End
            End
            Iteration += 1
        End
    End
```

Figure 2: Pseudo code

Here fitness function is calculated with the help of Support Vector Machine, where RBF kernel is used for classification since Gaussian function makes a good cluster pattern distribution in N dimensional space vector [Text Feature Selection using Particle Swarm Optimization Algorithm]. The reason for using Gaussian function is because it obtains fine granularity over the particle centroid from each point of the data Picton (2000). When compared to back propagation method RBF provides high efficiency over non stationary inputs. On considering from previous studies, it is seen that implantation on BBC dataset have not considered using RBF network. For performing this process in an efficient way a package called LIBSVM is used, as LIBSVM can work with multi classification and regression related problems. Input of the LIBSVM comprises of basically three components and it should be represented in the below format.

```
<class> <FeatureID1>:<value1> <FeatureID2>:<value2> ...
.
.
.
```

Figure 3: Data Representation

Where Class is the class to which the particular line belongs to, FeatureID is the ID of each feature and <value>is its IF-IDF values.

4 Implementation

Data from different category folders are combined together to form a single CSV file with the encoding called ISO-8859-1 by Python programing language. One all the file are appended to a dataframe, it leaves a final document consisting of 2225 rows 2 columns. Here the fields are Topic, Data. Where topic represents the Class and the Data are raw text files. Later each row in the data should undergo cleaning process, where the first stage is converting all the text to lowercase then tokenizing it. Next stage in cleaning is removing unwanted punctuations, numbers, symbols and blank spaces from the text. Next step in cleaning is stemming of the words. This is where the available text from the token is taken out and the true derived form of word is replaced to it. Say example having is derived from have. Final stage of cleaning is to remove stopwords, NLTK has inbuilt corpus for stopwords with the help of this all the unwanted words are removed from the list.

```
In [10]: len(filtere_sentence)
Out[10]: 1548
```

Figure 4: Total Number of Features

Final entire corpus consists of 1548 features which belongs to all 5 classes. Similarly same procedure is followed for each classes by filtering out data column which belongs to topic Business. This resulting in creating a separate corpus for each class. This method helps a lot in calculation of TF-IDF which is the key essential data conversion factor to this entire research. From the main corpus filtere sentence, filter out the unique words, in other words remove the duplicates. These are the feature which the entire TF-IDF value to be calculated on which is of total 695, so upload it to the data frame which contains topic and data. SO the final document consists of topic, data and features as shown below comprising of 2225 rows * 983 columns fig (5)

Inorder to understand each corpus which is created during the process, a word cloud is generated, this gives a pictorial view of the corpus belonging to each category fig (6).

By using sklearn library the value of each text file is calculated and appended to its corresponding place feature value in the dataframe and saved to CSV format. This datasets are finally used for the PSOSVM process.

In [43]: data_df

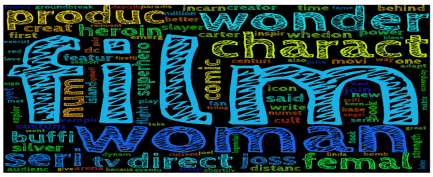
Out[43]:

	topic	data	abl	abolish	absolut	absorb	abus	access	accord	account	...	worri	wors	worst	write	wrote	xbox	
0	business	Ad sales boost Time Warner profit\r\n\r\nQuart...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	I
1	business	Dollar gains on Greenspan speech\r\n\r\nThe do...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	I
2	business	Yukos unit buyer faces loan claim\r\n\r\nThe o...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	I
3	business	High fuel prices hit BA's profits\r\n\r\nBriti...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	I
4	business	Pernod takeover talk lifts Domecq\r\n\r\nShare...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	I
5	business	Japan narrowly escapes	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	I

Figure 5: Created Data Frame

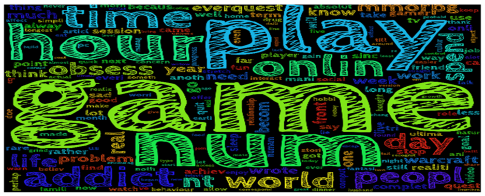
Business WordCloud

Politic WordCloud



Entertainment WordCloud

Sports WordCloud



Technology WordCloud

Figure 6: Created Data Frame

Before starting the process of PSOSVM, the experiment must be conducted in a regular parallel PSO technique Gaikwad and Patwardhan (2014). At every iteration random particles have to be initiated and then it has to be assigned to the data and not when every data is updated. So when new data is updated, the particle which are not assigned to the data are reinitialized and assigned to the particle. The fitness value is calculated with the help of SVM of RBF kernel SVC of each particle and the calculated value is assigned to the feature number of the subsets. For calculating the effectiveness of optimization being produced here we compare TFIDF value of each iteration with the term class on the training dataset. Inorder to find the exact pattern matching the document class, all the features are considered except the features which provides zero TFIDF value. Then 10-fold cross validation accuracy is calculated using LIBSVM with SVM RBF kernel type. Similar process is carried out for SPSO, which is a simplified location updating of BPSO.

5 Evaluation

Inorder to test the PSO and to do some basic calibration, the algorithm is modified by reducing the parameters and the instead of svm as a fitness function the parabolic equation is used. So the aim of the process is to find a global solution on a single dimensional vector plane. This will provide a random distribution of particles on the plane of parabola. During the first iteration particles are scattered on the parabolic surface and assigned to the fitness function data. After every iteration gbest value is calculates and moved to a new position. Finally all the particles move towards the global solution from local solution, this movement can be seen in the below image.

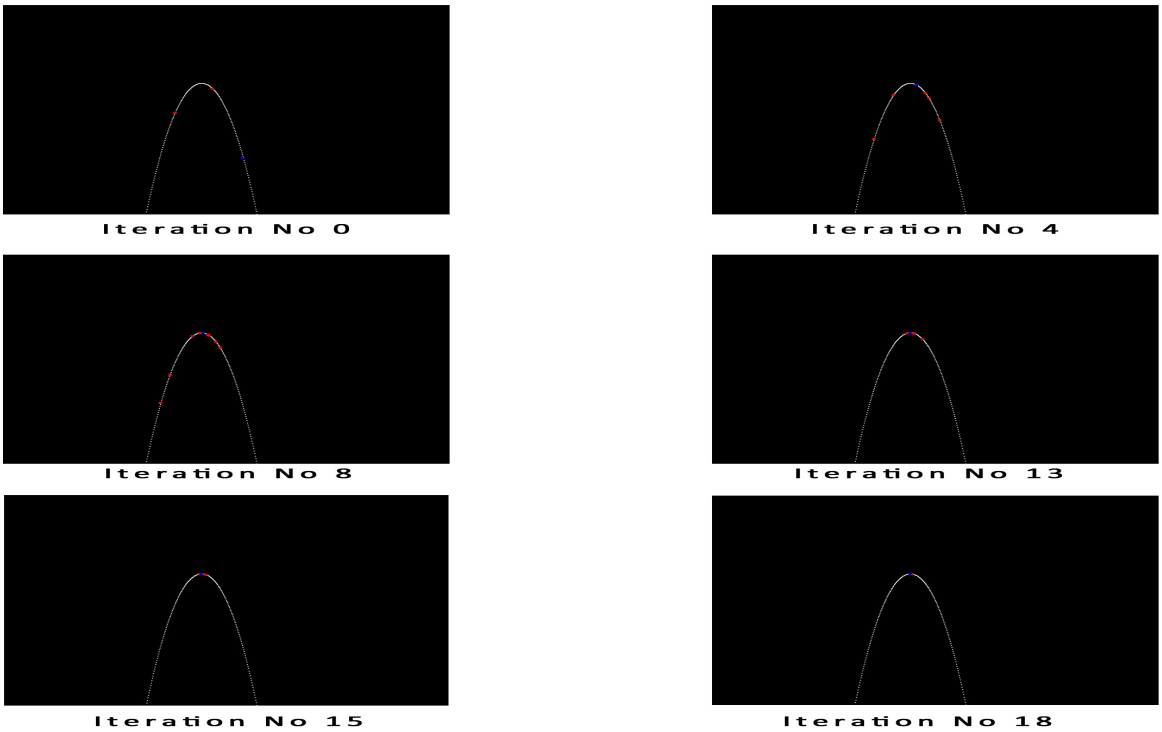


Figure 7: Testing

Its error is the error value of each particle obtained from the cross validation.

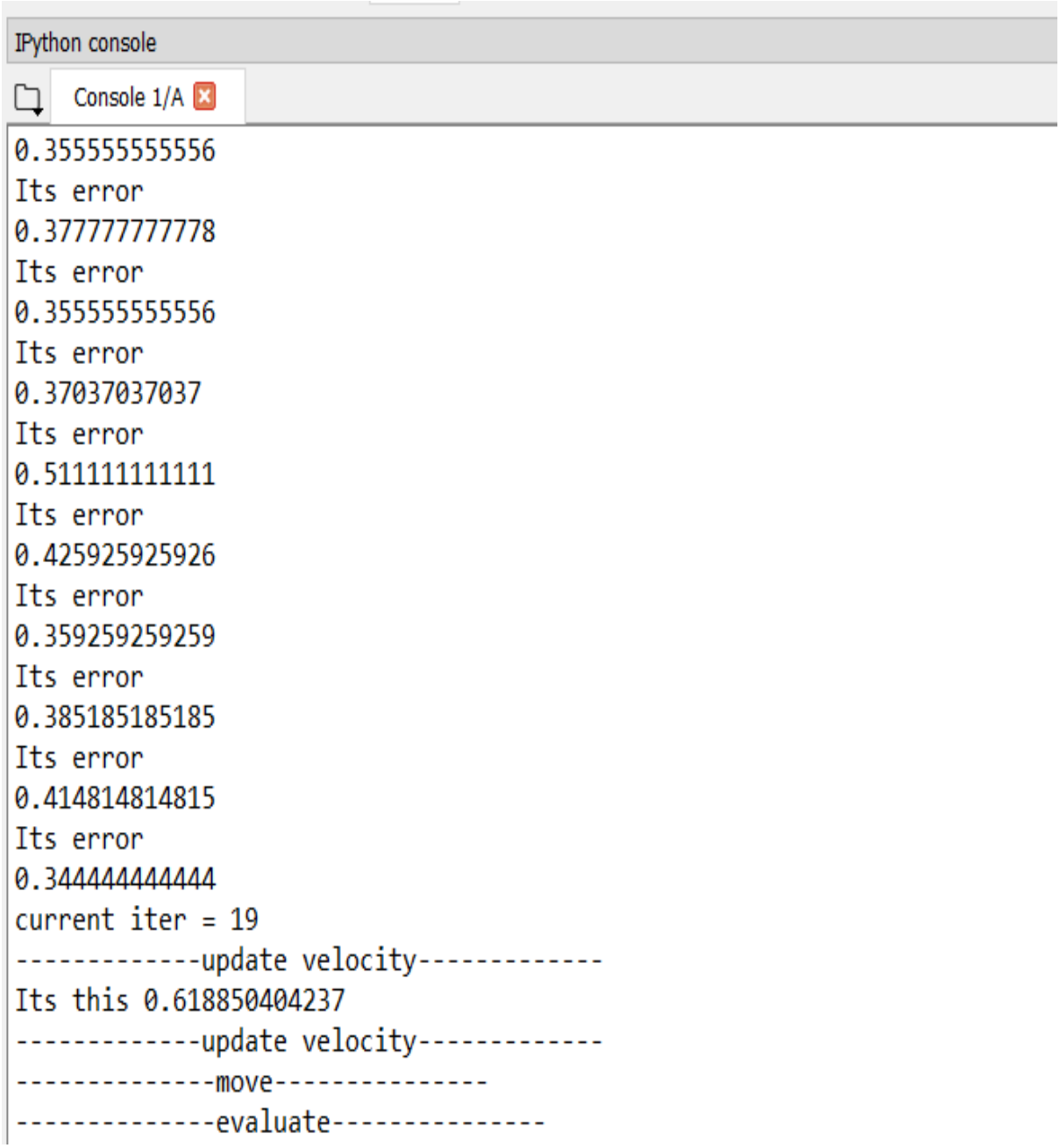


Figure 8: Error calculation

If a global solution is obtained velocity of the particle is updated and the move function gets activated, requesting the particle to move towards the global solution. This process is carried out at every update in global value occurs.

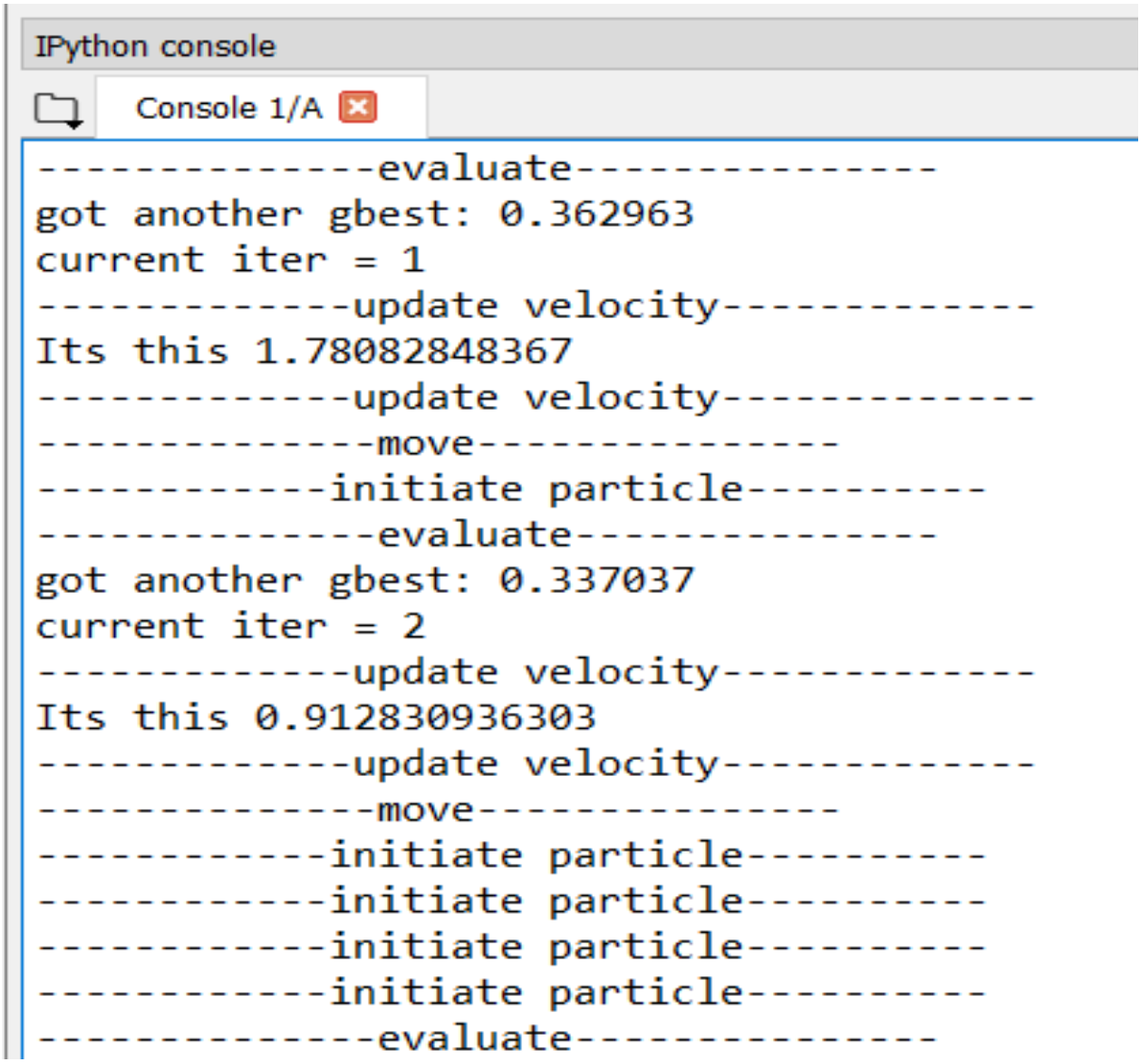


Figure 9: Working Mechanism

Number of Particles	Number of iterations	Error output
15	50	2.7
35	100	1.98
50	150	0.85

Figure 10: PSO Analysis

Number of Particles	Number of iterations	Error output of BPSO	Error output of SPSO
15	50	2.7	4.3
35	100	1.98	1.45
50	150	0.85	0.32

Figure 11: PSO and SPSO Analysis

5.1 Discussion

In this over all research process there were challenging parts which couldnt be cover come such as lack of memory on the machine, computational capability. As we can see that due to these hurdles generation of pictorial representation was not possible when processing the whole data due to higher dimensionality computational capability. Based on the results which are generated we can say if these problems are overcome, a pictorial representation and higher extent iteration of the process would have provided a much granular results.

6 Conclusion and Future Work

Studied conducted for this research where based on Particle Swarm Optimization feature selection Algorithm and provided support from the experiment stated that this optimization algorithm in combination with SVM provide is much efficient way of handling higher dimensional datasets. Not only in terms of high dimensionality features also supports more when there are more number of classification. Here fitness function provides an important key feature, so that a changes in the fitness function will have a major impact on the PSO performance. From the research performed we can conclude that PSO is a powerful search based algorithm and can still be improvised by tuning the input parameters. So by using PSO with LIBSVM higher dimensional multivariate problem can be solved in granular level.

Future work will be based on resolving the problems faced during this process which includes unpacking the predicted dataset from the classified solution and to implement mapreduce on SPARK. Since it a higher dimensional feature reduction type problem solving, usage of hadoop platform will increase the computational capability and parallel processing of data to reduce the processing time. As PSO is basically a parallel processing, by using hadoop efficient can be increased. Not only that more complicated analysis can be conducted over it. Also in future a fine tuning analysis of C-SVC according to the SVM model can yield a granular level classification, which will be very much helpful in identifying the classification problem.

Acknowledgements

In performing this research, I am very much thankful for the endless guidance and help provided by Dr. Jason Roche and National College of Ireland. I also want to be thankful for my friends and family on helping to do this research.

References

Al-Ani, A. (7). Ant colony optimization for feature subset selection, *International Science Index, Computer and Information Engineering* **1**(4).

B, E. (n.d.). *Cluster Analysis* .

Bird, S., Klein, E. and Loper, E. (2009). *Natural Language Processing with Python*, 1st edn, O'Reilly Media, Inc.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines, *ACM Trans. Intell. Syst. Technol.* **2**(3): 27:1–27:27.
URL: *http://doi.acm.org/10.1145/1961189.1961199*

Cui, X., Potok, T. E. and Palathingal, P. (2005). Document clustering using particle swarm optimization, *Proceedings 2005 IEEE Swarm Intelligence Symposium, 2005. SIS 2005.*, pp. 185–191.

Dong and Hownet, Q. D. (n.d.). Natural language processing.
URL: *http://www.keenage.com*

EVERS, G. I. (2009). An automatic regrouping mechanism to deal with stagnation in particle swarm optimization.
URL: *http://www.georgeevers.org/thesis.pdf*

Gaikwad, K. S. and Patwardhan, M. S. (2014). Tweets clustering: Adaptive pso, *2014 Annual IEEE India Conference (INDICON)*, pp. 1–6.

How, B. C. and Narayanan, K. (2004). An empirical study of feature selection for text categorization based on term weightage, *Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on*, pp. 599–602.

Hu, X. (n.d.). Particle swarm optimization.
URL: *http://www.swarmintelligence.org/index.php*

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999). Data clustering: A review, *ACM Comput. Surv.* **31**(3): 264–323.
URL: *http://doi.acm.org/10.1145/331499.331504*

Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization, *Neural Networks, 1995. Proceedings., IEEE International Conference on*, Vol. 4, pp. 1942–1948 vol.4.

Lai, C. C. and Wu, C. H. (2007). Particle swarm optimization-aided feature selection for spam email classification, *Second International Conference on Innovative Computing, Informatio and Control (ICICIC 2007)*, pp. 165–165.

l.G. Hao and l.O. Wang (2013). The method of text feature selection based on lsi and ga, *Information Science* .

Lin, S.-W., Ying, K.-C., Chen, S.-C. and Lee, Z.-J. (2008). Particle swarm optimization for parameter determination and feature selection of support vector machines, *Expert Syst. Appl.* **35**(4): 1817–1824.
URL: *http://dx.doi.org/10.1016/j.eswa.2007.08.088*

Picton, P. (2000). Neural networks, *Palgrave* .

Prusa, J. D., Khoshgoftaar, T. M. and Napolitano, A. (2015). Using feature selection in combination with ensemble learning techniques to improve tweet sentiment classification performance, *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 186–193.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval, *Inf. Process. Manage.* **24**(5): 513–523.
URL: *http://dx.doi.org/10.1016/0306-4573(88)90021-0*

Salton, G. and McGill, M. (1983). Introduction to modem information retrieval, *McGraw-Hill* .

StuartReid (n.d.). Portfolio optimization using particle swarm optimization.
URL: *http://web.archive.org/web/20080207010024*

Sun, F. and Sun, M. (2005). A new transductive support vector machine approach to text categorization, *2005 International Conference on Natural Language Processing and Knowledge Engineering*, pp. 631–635.

VS, T. J. A. and P, T. (2013). Paradigms for realizing machine learning algorithms, *Bigdata* .

Wang, X., Yang, J., Teng, X., Xia, W. and Jensen, R. (2007). Feature selection based on rough sets and particle swarm optimization, *Pattern Recogn. Lett.* **28**(4): 459–471.
URL: *http://dx.doi.org/10.1016/j.patrec.2006.09.003*

Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization, *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 412–420.
URL: *http://dl.acm.org/citation.cfm?id=645526.657137*

Zhang, X., Wang, T. and Fang, J. (2014). A node localization approach using particle swarm optimization in wireless sensor networks, *2014 International Conference on Identification, Information and Knowledge in the Internet of Things*, pp. 84–87.