Review of Computer Engineering Research
2014 Vol.1, No.1, pp.19-30
ISSN(e): 2410-9142
ISSN(p): 2412-4281

WEB PAGES CATEGORIZATION BASED ON CLASSIFICATION & OUTLIER ANALYSIS THROUGH FSVM

Geeta R.B.*[1] --- Shobha R.B.[2] --- Shashikumar G.Totad[3] --- Prasad Reddy PVGD[4]

[1]*Department of Information Technology, GMR Institute of Technology, RAJAM, AP, India*

[2]*Department of Electronics and Communications, Basaweshwar Engineering College, Bagalkot, India*

[3]*Department of Computer Science, GMR Institute of Technology, RAJAM, AP, India*

[4]*Department of CS & SE, Andhra University, Vizag, AP, India*

ABSTRACT

The performance of Support Vector Machine is higher than traditional algorithms. The training process of SVM is sensitive to the outliers in the training set. Here in this Paper, a new approach called, Web Pages Categorization based on Classification and Outlier Analysis (WPC-COA), is proposed that uses a polynomial Kernel function to map web page tuples to high dimensional feature space.

Keywords: Support vector machine, Outliers, Categorization, Log file, Kernel parameters, Web page.

Contribution/ Originality

This study uses a new methodology which helps in mapping web page tuples with various attributes such as frequency, time spent on each page, in-degree, out-degree and level of a web page to high dimensional feature space. The paper's primary contribution is to categorize web pages based on classification and outlier analysis using Polynomial Kernel function.

1. INTRODUCTION

Intelligence has been well-defined in several ways including logical reasoning, abstract thinking power, better understanding, self-awareness, interacting with others, learning by experiences, being emotional, retaining, scheduling and better problem resolving capabilities. Intelligence is extensively observed in humans, but also perceived in animals and plants. Artificial Intelligence is the modeling of intelligence in machines. Machine learning is a branch of Artificial Intelligence which used in building systems that can learn from data. This learning is further used in building systems that can learn from test data sets.

In machine learning, Support Vector Machine (SVM) is a supervised learning simulation integrated with learning algorithms that help in analysis, and distinguish patterns used for

† Corresponding author

classification and regression analysis. The basic SVM model accepts a set of input data and predicts which of the two possible classes the output belongs, making it a non-probabilistic binary linear classifier. For a given training data set, the set is marked as belonging to one of the two categories; SVM constructs a model that earmarks a given data set into one category or the other. SVM is a prominent approach for classification problems [1] and nonlinear function approximation problems in the fields of machine learning and pattern recognition. It provides accuracy measure of fuzzy classification for solving real-world problems [2]. Section 2 explains Support Vector Machine (SVM) and discusses about Fuzzy Support Vector Machine (FSVM) and Fuzzy C-Means Clustering (FCM). Section 3 presents FSVM's Web Pages Categorizer based on Classification and Outlier Analysis (WPC-COA). Section 4 shows experimental results.

2. SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine is a technique for classification of both linear and nonlinear data. Web page tuple is a record of one or more attributes such as in-degree, out-degree, level, frequency and utility etc..of a page. Let 'W' be the web page (data) set, collection of web pages, denoted as{(P1, B1), (P2, B2)….. (Pn, Bn)], where Pi is a tuple of page 'i' with associated class label Bi. Each Bi can take one of two values either positive class (+1/Yes) or negative class (-1/ No) as shown in Fig. 1. For a web page set with 2 attributes/dimensions, infinite number of separators can be drawn. This can be generalized to n dimensions/attributes. This optimal separator is known as hyperplane [3]. Hyperplane with a larger margin is more accurate than with smaller margin.

Fig-1. Web pages separator

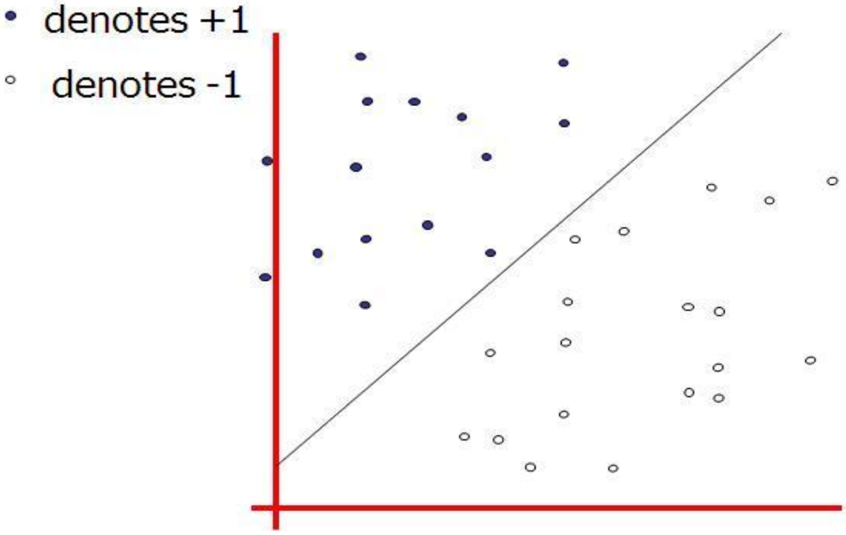
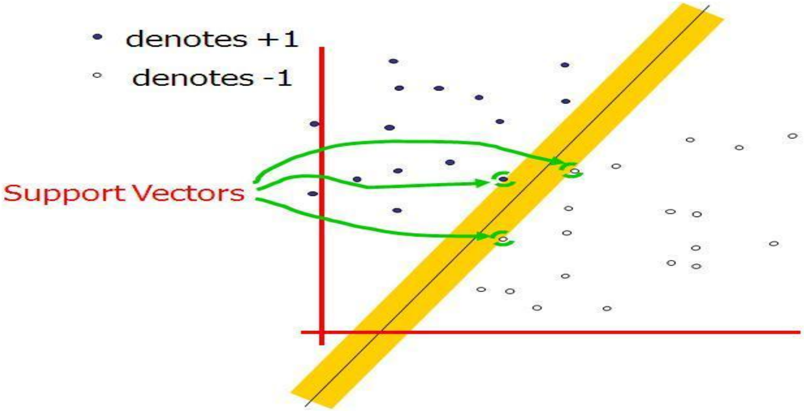


Fig-2. Web Page as Support Vectors



A combination of few input data points, known as support vectors [3], can be an optimal solution as shown in Fig. 2.This SVM method of classification works fine when web page tuples are linear. Such a SVM is called linear SVM. For cases, where it is not possible to have a hyperplane as straight line or when web page tuples are nonlinear, extended SVM can be used. In such a case, input web page points are mapped into high dimensional feature space using nonlinear mapping, resulting in quadratic optimization problem. This problem can be solved using linear SVM. Optimal hyperplane in high dimensional feature space corresponds to nonlinear separating hyper surface in the original space. Linear and nonlinear SVMs are sensitive to noise or outliers. This is one of the main limitations of the standard SVM. In most of the real world classification problems, some training points do not belong to any of the predefined classes exactly, whenever noise or outlier exists. For example, one training point may belong 90% to one class and be 10% to none of the classes, and it may also belong 30% to one class and 70% none of the classes. SVM is a powerful tool for solving classification problems. Comparison [4] between traditional learning and SVM has shown that SVM provides better results than traditional learning in many applications.

2.1. Fuzzy Support Vector Machine (FSVM)

SVM draws optimal hard boundary hyperplane between two classes. It finds optimal hyperpalne using dot product functions, called Kernels. Each input web page may not completely belong to one of the two classes (+ve, -ve). This limitation can be overcome by Fuzzy Support Vector Machine (FSVM). Each input web page is assigned a fuzzy membership degree, so that input web page can make different contributions to the decision feature surface learning.

3. FUZZY C-MEANS CLUSTERING

Fuzzy C-means (FCM) is a clustering method. It allows one web page to belong to two or more clusters. This method proposed by Dunn in 1973, was modified in 1981 by Bezdek. FCM's aim is to minimize the following objective function

$$Fp = \sum_{i=1}^{N} \sum_{i=1}^{C} m_{ij}{}^{p} \; \| \, \| x_i - c_j \| \, \| \; ^2$$

$$1 \leq p < \infty$$

where 'p' is a number greater than 1, m$_{ij}$ is the degree of membership of web page x_i in cluster *j*, x_i is the *i*th web page with 'd' attributes, c_j is d-dimension center of cluster j, the center c$_j$ and any measured data's similarity is expressed using any norm ||*||is any norm expressing the similarity between any measured data and the center c$_j$.

Fuzzy partitioning is carried out through iterative optimization of the objective function shown above, by updating membership m$_{ij}$ and the cluster centers c_j as

$$m_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\| x_i - c_j \| / \| x_i - c_k \| \right)^{2 \div (m-1)}}$$

$$c_j = \frac{\sum_{i=1}^{N} m_{ij}{}^{p} \, . x_i}{\sum_{i=1}^{N} m_{ij}{}^{p}}$$

This process is iterated till

$$max_{ij}\{ | m_{ij} \; ^{(k+1)} - m_{ij} \; ^{(k)} | < \varphi \quad ,$$

where φ value lies between 0 and 1 and is called as termination criterion, whereas *k* is the iteration number. This procedure ultimately converges to a local minimum.

4. FSVM'S WEB PAGES CATEGORIZER BASED ON CLASSIFICATION AND OUTLIER ANALYSIS ALGORITHM

In FCM, the sum of memberships of web pages across classes is one. FCM uses probabilistic constraint. Krishnapuram and Keller relaxedthis constraint and proposed a possibilistic approach to clustering (PCM). Kernel Possibilistic C-means (KPCM) substitutes kernel-induced distance metric for the Euclidean distance [5]. KPCM algorithms avoid clusters coincidence and are less sensitive to outliers.

Consider website with n pages, n>0. Each page is associated with attributes such as in-degree (the number of web pages pointing to this page), out-degree (the number of pages to which this page is pointing), level, frequency and time spent on each page. Assuming there are 'm' web page tuples, 80% of 'm' can be used for training and remaining 20% of 'm' can be used for testing. Polynomial Kernel is used to get a hyperplane.

$K\ (Xi, Xj) = (Xi \cdot Xj + b)^d$

Where 'd' is a degree of the polynomial function, and 'b' is a constant.

Given number of clusters 'C', degree of the polynomial function'd' and constant 'b', dot product of Xi and Xj can be obtained. If K (Xi, Xj) <0 then the tuple belongs to negative class otherwise it belongs to positive class. Outliers are not erroneous. Outliers are the tuples that do not comply with general behavior of the data. These are tuples which are far away from the other tuples in the given data space. Certain web page details are inconsistent with the remaining set of web page tuples, for example, a wrong entry of web page in-degree as -1. Noise can be of two types; attribute noise and class noise. Missing value of particular attribute (level of web page is missing) is an example of attribute noise. Class noise is of two types, namely contradictory examples and misclassifications. Example of contradictory noise, include the web page sets appearing more than once and being labeled with different classes. Examples of misclassification include two different classes with same symptoms or features. Here in this chapter, classification problems with outliers or misclassification noise are dealt with. Outliers and noise are assigned to lower membership degrees. Two farthest pair of clusters are classified to form a binary classification problem. This process degrades the effects of noise or outlier on the decision function. When the farthest pair of clusters is adopted, and outliers are degraded, FSVM is applied to obtain final classification results.

Consider a web site of 'n' web pages. The structure of website provides information about in-degree, out-degree, level, frequency of a web page and server log file helps in getting information about time spent on each web page in a specific period of time. These can be attributes of a web page. Get n pairs of web page tuples {Xi , Yi, Si} for a binary classification problem, where Xi $\in R^n$are the input web pages, Yi$\in$ {−1, +1] are the corresponding binary class labels, and Si$\in$ (0, 1] is the fuzzy membership degree of Xi belonging to Y. Let X[i] denote a web page with five attributes such as frequency, time spent, in-degree, out-degree and level of a web page. Sixth attribute is used as class label, which labeled as positive class (+1) or negative class (-1). 80% of log file data is used as training web pages. 20% log file data is used as testing data.

4.1. Algorithm WPC-COA (Server Log File, Website, Kernel Parameters, No. of Clusters)

1. Get frequency and time spent on each page from server log file.
2. Get in-degree, out-degree and level of each web page from website structure.
3. For each training web page tuple, sixth attribute of each web page is used as class label which is assigned a label either positive class (1) or negative class (-1).
4. Get n pairs of web page tuples {Xi , Yi, Si} for a binary classification problem, where Xi $\in R^n$are the input web pages, Yi$\in$ {−1, +1} are the corresponding binary class labels, and Si$\in$ (0, 1] is the fuzzy membership degree of Xi belonging to Y.

5. Read number of clusters, 'C' and termination parameter 'φ'.
6. Read 'b' and degree of polynomial function 'd'.
7. Initialize cluster centroids Xj, j = 1, 2. . . C.
8. Map training web page tuples into high dimensional feature space using FCM and cluster them into positive class (P) or negative class (N).
9. Decide whether pair, (Xi, Xj) of web page belongs to P or negative class N and to which cluster, using Kernel polynomial function.

   $K(Xi\ , Xj) = (Xi\ \cdot Xj+ b)^{d}$. d is a degree of the polynomial function, and b is an constant.
10. Let number of clusters in class P be Z+ and number of clusters in class N be Z-.
11. Search for farthest pair of clusters, in which one cluster belongs to Z+ and the other belongs to Z-.
12. Form a new training set with membership degrees, combining farthest pair of clusters and obtain a nonlinear classifier using FSVM with the parameters αi and b.
13. Stop

5. EXPERIMENTAL RESULTS

WPC-COA algorithm is implemented using java. Statistics collected from log file is stored in Excel with fields/attributes, frequency (f1)-number of times a web page is accessed, utility (f2)-time spent on a web page, number of backward links (f3), level (f4) of a web page. Field (f5) is used to randomly assign class label (positive +1, negative -1). Among 'n' number of log file records, 80% of web page tuples form training web page data set and remaining 20% of web page tuples are used as testing web page data set. Fig 3 to Fig. 9 shows output snapshots. Fig. 3 provides the user interface. When Browse button is clicked, it prompts for the location of file which contains web pages information. In Fig. 4 user selects the required file containing information about details of log file, with attributes such as frequency (f1), utility (f2), number of backlinks (f3), level (f4) and f5 with class label(+1 or -1). Fig. 5 shows testing data set. Fig. 6 and Fig. 7 prompt user for constant b and number of clusters respectively. Fig. 8 displays different clusters with positive and negative classes. Fig. 9 outputs distance between various clusters such as cluster distance between positive cluster 1 and negative cluster 1, positive cluster 1 and negative cluster 2, positive cluster 2 and negative cluster 1, positive cluster 2 and negative cluster 2.

Fig-3. GUI for Dataset selection

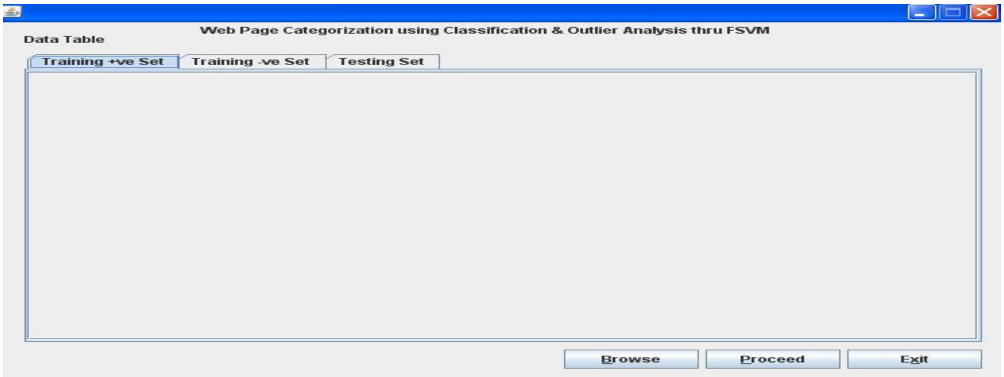
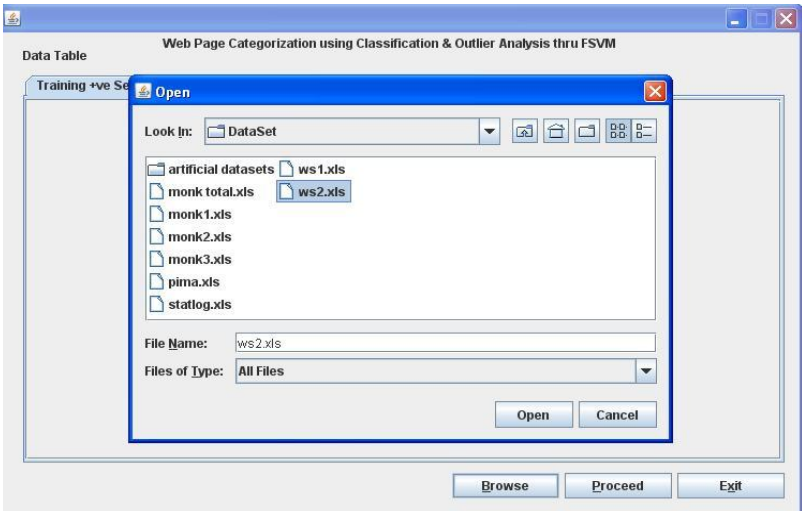

Web Page Categorization using Classification & Outlier Analysis thru FSVM
Data Table
Training +ve Set
Training -ve Set
Testing Set
Browse
Proceed
Exit

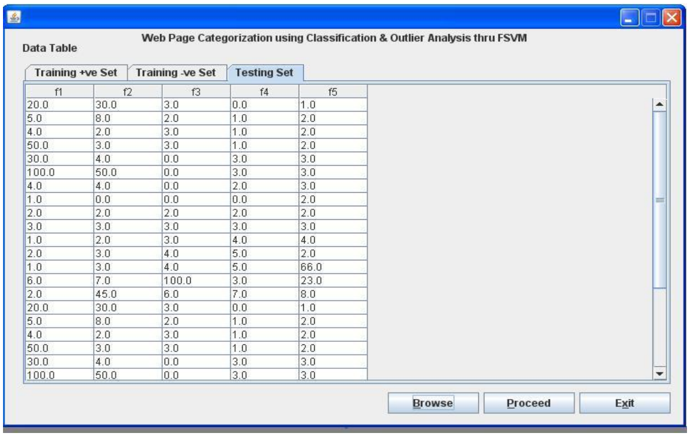
Fig-4. select the dataset



File Name: ws2.xls

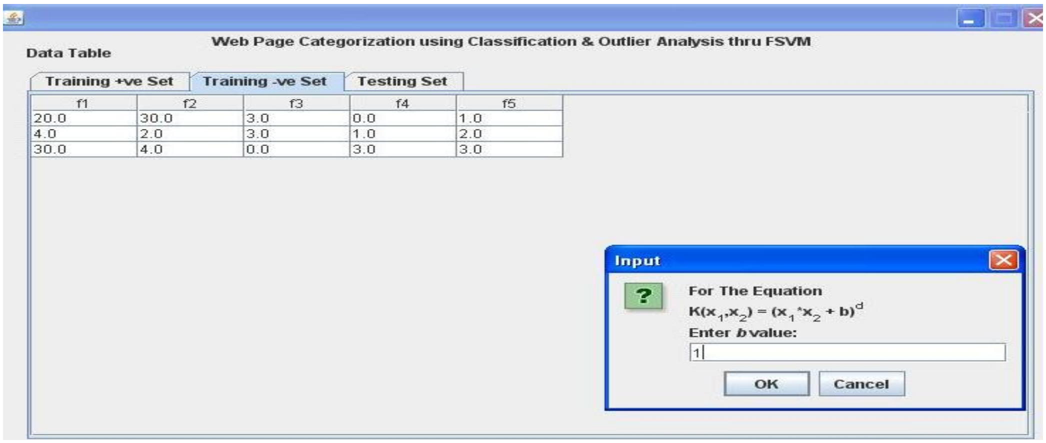
Files of Type: All Files

Fig-5. Testing Dataset


Web Page Categorization using Classification & Outlier Analysis thru FSVM
Data Table
Training +ve Set
Training -ve Set
Testing Set
Browse
Proceed
Exit

f1	f2	f3	f4	f5
20.0	30.0	3.0	0.0	1.0
5.0	8.0	2.0	1.0	2.0
4.0	2.0	3.0	1.0	2.0
50.0	3.0	3.0	1.0	2.0
30.0	4.0	0.0	3.0	3.0
100.0	50.0	0.0	3.0	3.0
4.0	4.0	0.0	2.0	3.0
1.0	0.0	0.0	0.0	2.0
2.0	2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0	3.0
1.0	2.0	3.0	4.0	4.0
2.0	3.0	4.0	5.0	2.0
1.0	3.0	4.0	5.0	66.0
6.0	7.0	100.0	3.0	23.0
2.0	45.0	6.0	7.0	8.0
20.0	30.0	3.0	0.0	1.0
5.0	8.0	2.0	1.0	2.0
4.0	2.0	3.0	1.0	2.0
50.0	3.0	3.0	1.0	2.0
30.0	4.0	0.0	3.0	3.0
100.0	50.0	0.0	3.0	3.0

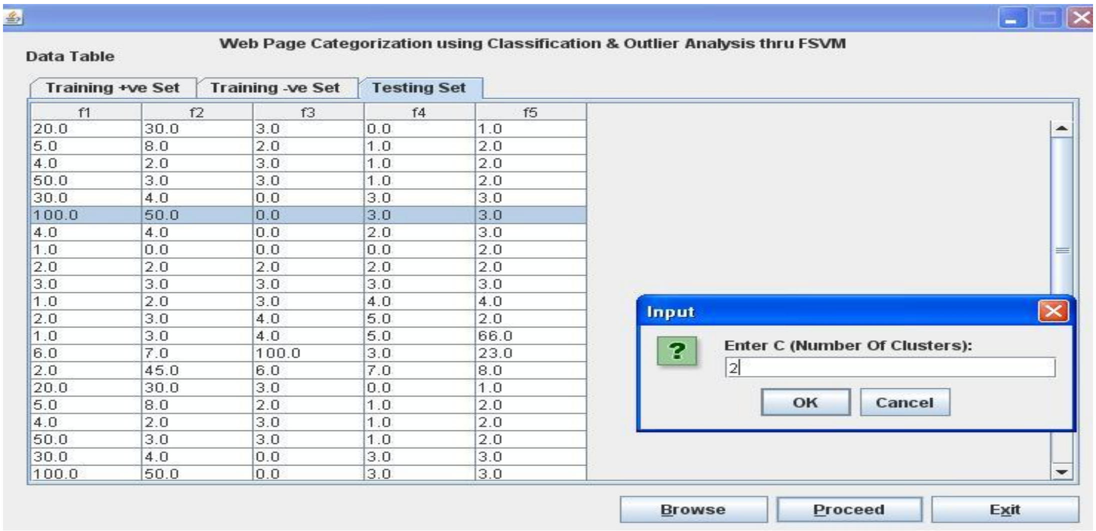
Fig-6. Browse the dataset Enter b value.


Web Page Categorization using Classification & Outlier Analysis thru FSVM
Data Table
Training +ve Set
Training -ve Set
Testing Set
Input
OK
Cancel

f1	f2	f3	f4	f5
20.0	30.0	3.0	0.0	1.0
4.0	2.0	3.0	1.0	2.0
30.0	4.0	0.0	3.0	3.0

For The Equation
$K(x_1,x_2) = (x_1 * x_2 + b)^d$
Enter *b* value:
1

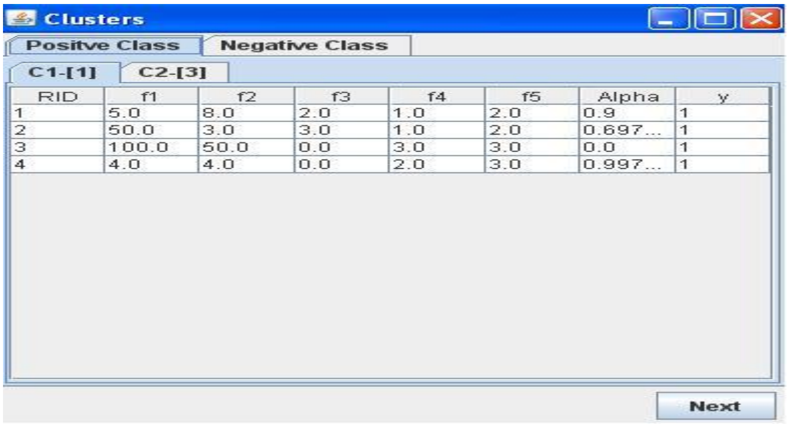
Fig-7. Enter C (number of Clusters)



f1	f2	f3	f4	f5
20.0	30.0	3.0	0.0	1.0
5.0	8.0	2.0	1.0	2.0
4.0	2.0	3.0	1.0	2.0
50.0	3.0	3.0	1.0	2.0
30.0	4.0	0.0	3.0	3.0
100.0	50.0	0.0	3.0	3.0
4.0	4.0	0.0	2.0	3.0
1.0	0.0	0.0	0.0	2.0
2.0	2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0	3.0
1.0	2.0	3.0	4.0	4.0
2.0	3.0	4.0	5.0	2.0
1.0	3.0	4.0	5.0	66.0
6.0	7.0	100.0	3.0	23.0
2.0	45.0	6.0	7.0	8.0
20.0	30.0	3.0	0.0	1.0
5.0	8.0	2.0	1.0	2.0
4.0	2.0	3.0	1.0	2.0
50.0	3.0	3.0	1.0	2.0
30.0	4.0	0.0	3.0	3.0
100.0	50.0	0.0	3.0	3.0

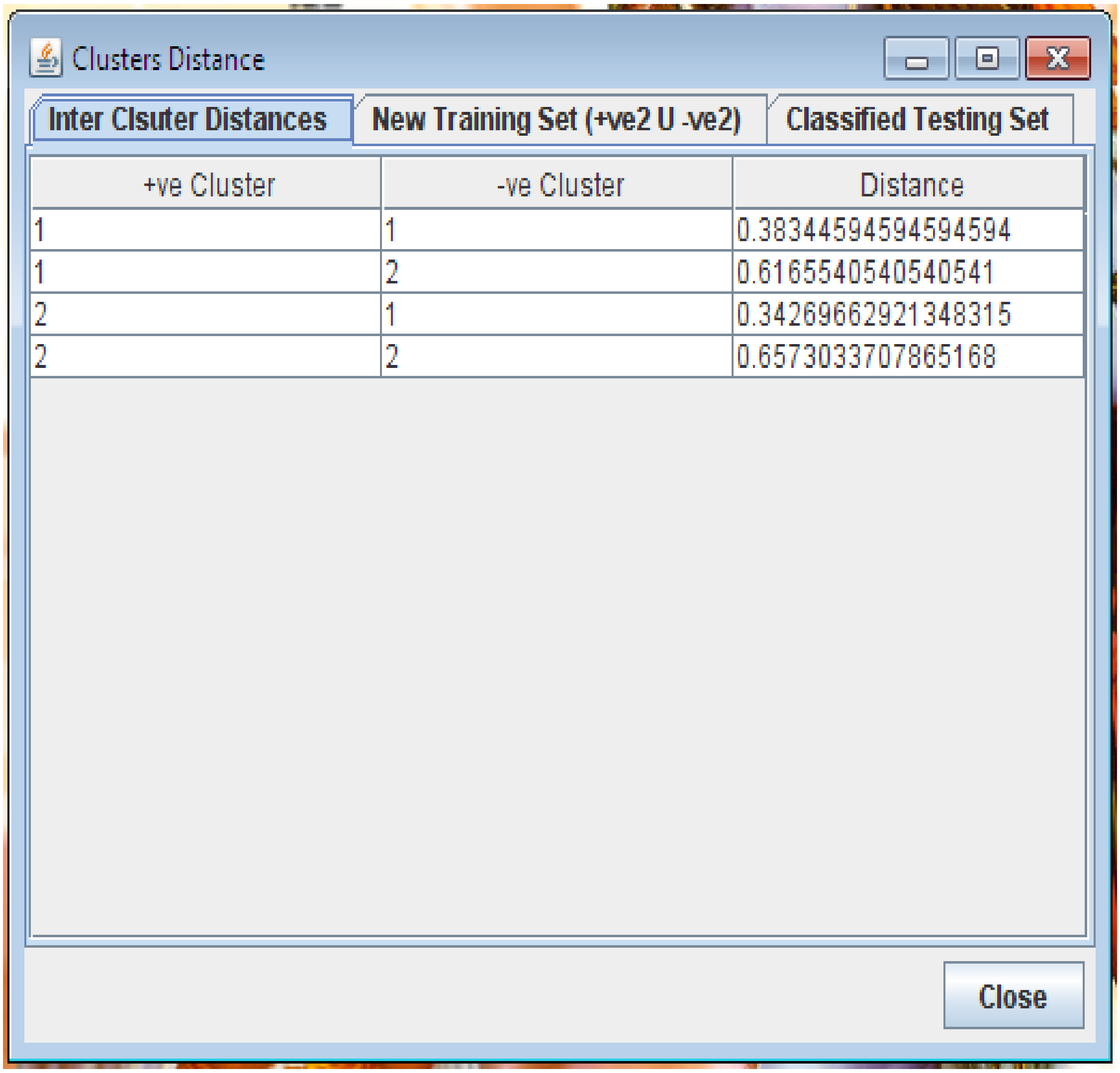
Enter C (Number Of Clusters):
2

Fig-8. Different clusters with positive and negative classes.



RID	f1	f2	f3	f4	f5	Alpha	y
1	5.0	8.0	2.0	1.0	2.0	0.9	1
2	50.0	3.0	3.0	1.0	2.0	0.697...	1
3	100.0	50.0	0.0	3.0	3.0	0.0	1
4	4.0	4.0	0.0	2.0	3.0	0.997...	1

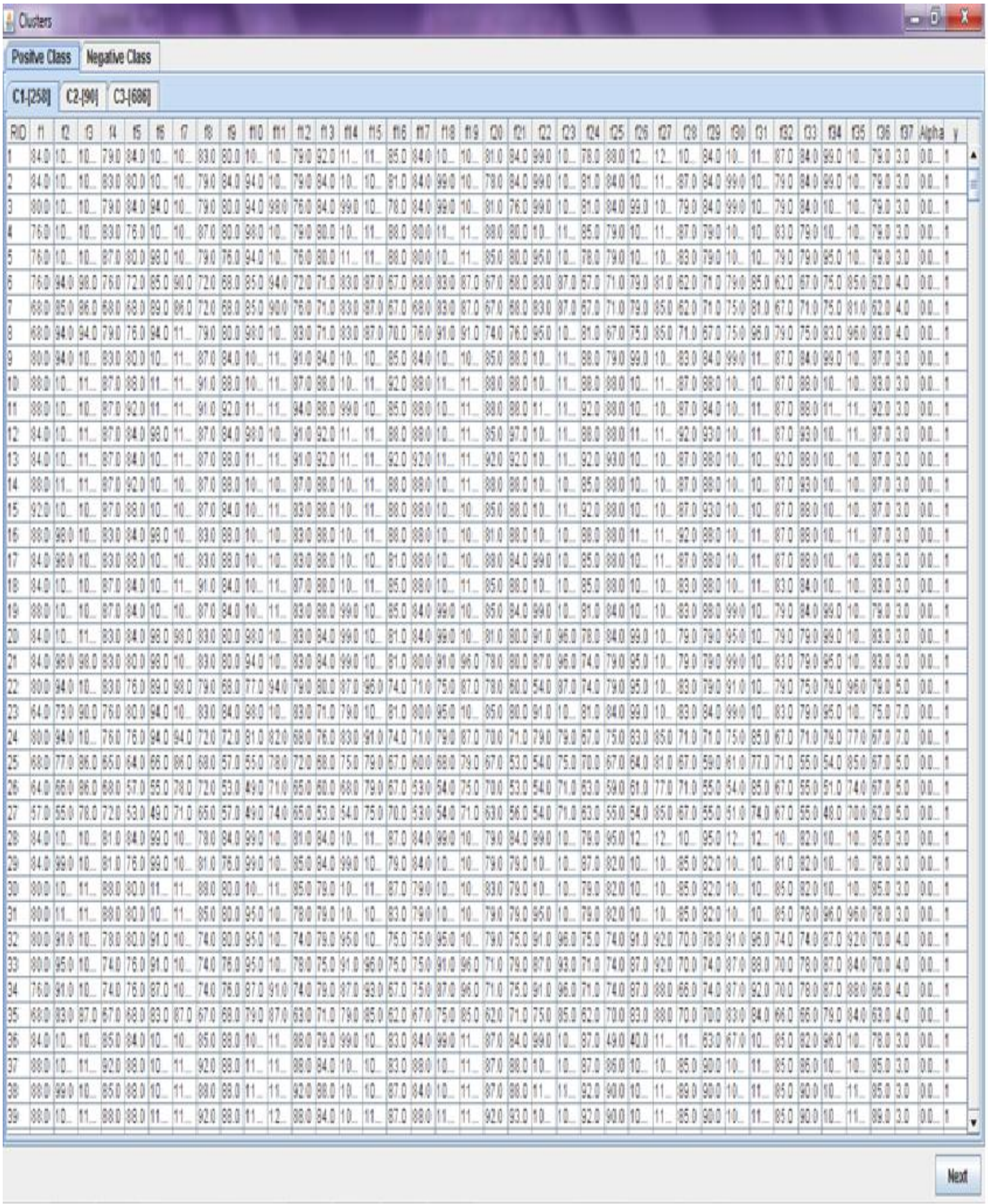
Fig-9. Clusters distance



+ve Cluster	-ve Cluster	Distance
1	1	0.38344594594594594
1	2	0.6165540540540541
2	1	0.3426966292134831 5
2	2	0.6573033707865168

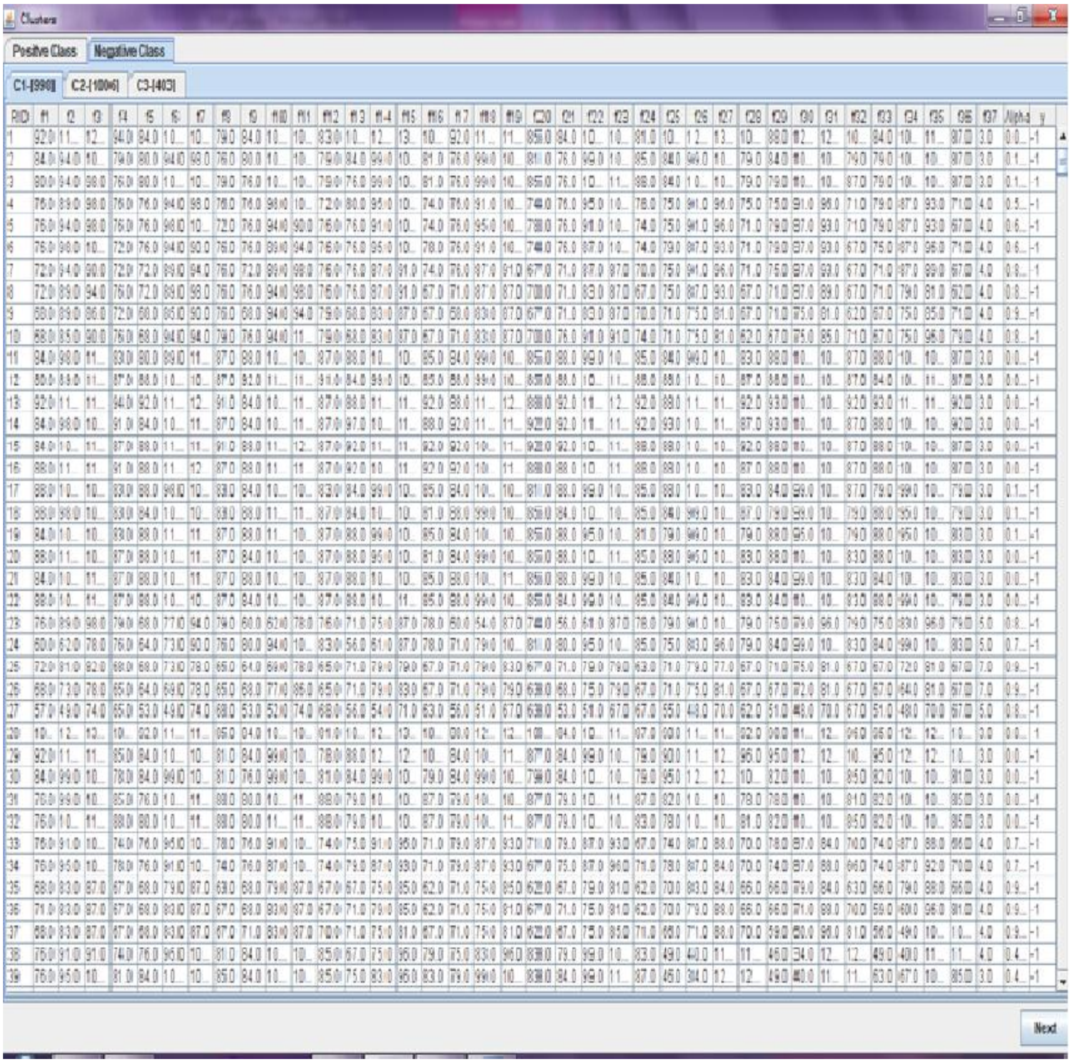
The file statlog contains 37 (f1, f2, f3….f7) attributes and 2218 records. Fig. 10 displays different clusters with positive classes. Fig. 11 displays different clusters with negative classes. Fig. 12 outputs distance between various clusters such as cluster distance between positive cluster 1 and negative cluster 1, positive cluster 1 and negative cluster 2, positive cluster 2 and negative cluster 1, positive cluster 2 and negative cluster 2, positive cluster 1 and negative cluster 3, positive cluster 2 and negative cluster 3, positive cluster 3 and negative cluster 3.

Fig-10. Different clusters with positive class



RID	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	f17	f18	f19	f20	f21	f22	f23	f24	f25	f26	f27	f28	f29	f30	f31	f32	f33	f34	f35	f36	f37	Alpha	y
1	84.0	10...	10...	79.0	84.0	10...	10...	83.0	80.0	10...	10...	79.0	92.0	11...	11...	85.0	84.0	10...	10...	81.0	84.0	99.0	10...	78.0	88.0	12...	12...	10...	84.0	10...	11...	87.0	84.0	99.0	10...	79.0	3.0	0.0...	1
2	84.0	10...	10...	83.0	80.0	10...	10...	79.0	84.0	94.0	10...	79.0	84.0	10...	10...	81.0	84.0	99.0	10...	78.0	84.0	99.0	10...	81.0	84.0	10...	11...	87.0	84.0	99.0	10...	79.0	84.0	99.0	10...	79.0	3.0	0.0...	1
3	80.0	10...	10...	79.0	84.0	94.0	10...	79.0	80.0	94.0	98.0	76.0	84.0	99.0	10...	78.0	84.0	99.0	10...	81.0	76.0	99.0	10...	81.0	84.0	99.0	10...	79.0	84.0	99.0	10...	79.0	84.0	10...	10...	79.0	3.0	0.0...	1
4	76.0	10...	10...	83.0	76.0	10...	10...	87.0	80.0	98.0	10...	79.0	80.0	10...	11...	88.0	80.0	11...	11...	88.0	80.0	10...	11...	85.0	79.0	10...	11...	87.0	79.0	10...	10...	83.0	79.0	10...	10...	79.0	3.0	0.0...	1
5	76.0	10...	10...	87.0	80.0	98.0	10...	79.0	76.0	94.0	10...	76.0	80.0	11...	11...	88.0	80.0	10...	11...	85.0	80.0	95.0	10...	78.0	79.0	10...	10...	83.0	79.0	10...	10...	79.0	79.0	95.0	10...	79.0	3.0	0.0...	1
6	76.0	94.0	98.0	76.0	72.0	85.0	90.0	72.0	68.0	85.0	94.0	72.0	71.0	83.0	87.0	67.0	68.0	83.0	87.0	67.0	68.0	83.0	87.0	67.0	71.0	79.0	81.0	62.0	71.0	79.0	85.0	62.0	67.0	75.0	85.0	62.0	4.0	0.0...	1
7	68.0	85.0	86.0	68.0	68.0	89.0	86.0	72.0	68.0	85.0	90.0	76.0	71.0	83.0	87.0	67.0	68.0	83.0	87.0	67.0	68.0	83.0	87.0	67.0	71.0	79.0	85.0	62.0	71.0	75.0	81.0	67.0	71.0	75.0	81.0	62.0	4.0	0.0...	1
8	68.0	94.0	94.0	79.0	76.0	94.0	11...	79.0	80.0	98.0	10...	83.0	71.0	83.0	87.0	70.0	76.0	91.0	91.0	74.0	76.0	95.0	10...	81.0	67.0	75.0	85.0	71.0	67.0	75.0	95.0	79.0	75.0	83.0	96.0	83.0	4.0	0.0...	1
9	80.0	94.0	10...	83.0	80.0	10...	11...	87.0	84.0	10...	11...	91.0	84.0	10...	10...	85.0	84.0	10...	10...	85.0	88.0	10...	11...	88.0	79.0	99.0	10...	83.0	84.0	99.0	11...	87.0	84.0	99.0	10...	87.0	3.0	0.0...	1
10	88.0	10...	11...	87.0	88.0	11...	11...	91.0	88.0	10...	11...	87.0	88.0	10...	11...	92.0	88.0	11...	11...	88.0	88.0	10...	11...	88.0	88.0	10...	11...	87.0	88.0	10...	10...	87.0	88.0	10...	10...	83.0	3.0	0.0...	1
11	88.0	10...	10...	87.0	92.0	11...	11...	91.0	92.0	11...	11...	94.0	88.0	99.0	10...	85.0	88.0	10...	11...	88.0	88.0	11...	11...	92.0	88.0	10...	10...	87.0	84.0	10...	11...	87.0	88.0	11...	11...	92.0	3.0	0.0...	1
12	84.0	10...	11...	87.0	84.0	98.0	11...	87.0	84.0	98.0	10...	91.0	92.0	11...	11...	88.0	88.0	10...	11...	85.0	97.0	10...	11...	88.0	88.0	11...	11...	92.0	93.0	10...	11...	87.0	93.0	10...	11...	87.0	3.0	0.0...	1
13	84.0	10...	11...	87.0	84.0	10...	11...	87.0	88.0	11...	11...	91.0	92.0	11...	11...	92.0	92.0	11...	11...	92.0	92.0	10...	11...	92.0	93.0	10...	10...	87.0	88.0	10...	10...	92.0	88.0	10...	10...	87.0	3.0	0.0...	1
14	88.0	11...	11...	87.0	92.0	10...	10...	87.0	88.0	10...	10...	87.0	88.0	10...	11...	88.0	88.0	10...	11...	88.0	88.0	10...	10...	85.0	88.0	10...	10...	87.0	88.0	10...	10...	87.0	93.0	10...	10...	87.0	3.0	0.0...	1
15	92.0	10...	10...	87.0	88.0	10...	10...	87.0	84.0	10...	11...	83.0	88.0	10...	11...	88.0	88.0	10...	10...	85.0	88.0	10...	11...	92.0	88.0	10...	10...	87.0	93.0	10...	10...	87.0	88.0	10...	10...	87.0	3.0	0.0...	1
16	88.0	98.0	10...	83.0	84.0	98.0	10...	83.0	88.0	10...	10...	83.0	88.0	10...	11...	88.0	88.0	10...	10...	81.0	88.0	10...	10...	88.0	88.0	11...	11...	92.0	88.0	10...	11...	87.0	88.0	10...	11...	87.0	3.0	0.0...	1
17	84.0	98.0	10...	83.0	88.0	10...	10...	83.0	88.0	10...	10...	83.0	88.0	10...	10...	81.0	88.0	10...	10...	88.0	84.0	99.0	10...	85.0	88.0	10...	11...	87.0	88.0	10...	11...	87.0	88.0	10...	10...	83.0	3.0	0.0...	1
18	84.0	10...	10...	87.0	84.0	10...	11...	91.0	84.0	10...	11...	87.0	88.0	10...	11...	85.0	88.0	10...	11...	85.0	88.0	10...	10...	85.0	88.0	10...	10...	83.0	88.0	10...	11...	83.0	84.0	10...	10...	83.0	3.0	0.0...	1
19	88.0	10...	10...	87.0	84.0	10...	10...	87.0	84.0	10...	11...	83.0	88.0	99.0	10...	85.0	84.0	99.0	10...	85.0	94.0	99.0	10...	81.0	84.0	10...	10...	83.0	88.0	99.0	10...	79.0	84.0	99.0	10...	79.0	3.0	0.0...	1
20	84.0	10...	11...	83.0	84.0	98.0	98.0	83.0	80.0	98.0	10...	83.0	84.0	99.0	10...	81.0	84.0	99.0	10...	81.0	80.0	91.0	96.0	78.0	84.0	99.0	10...	79.0	79.0	95.0	10...	79.0	79.0	99.0	10...	83.0	3.0	0.0...	1
21	84.0	98.0	98.0	83.0	80.0	98.0	10...	83.0	80.0	94.0	10...	83.0	84.0	99.0	10...	81.0	80.0	91.0	96.0	78.0	80.0	87.0	96.0	74.0	79.0	95.0	10...	79.0	79.0	99.0	10...	83.0	79.0	95.0	10...	83.0	3.0	0.0...	1
22	80.0	94.0	10...	83.0	76.0	89.0	98.0	79.0	68.0	77.0	94.0	79.0	80.0	87.0	96.0	74.0	71.0	75.0	87.0	78.0	80.0	54.0	87.0	74.0	79.0	95.0	10...	83.0	79.0	91.0	10...	79.0	75.0	79.0	96.0	79.0	5.0	0.0...	1
23	64.0	73.0	90.0	76.0	80.0	94.0	10...	83.0	84.0	98.0	10...	83.0	71.0	79.0	10...	81.0	80.0	95.0	10...	85.0	80.0	91.0	10...	81.0	84.0	99.0	10...	83.0	84.0	99.0	10...	83.0	79.0	95.0	10...	75.0	7.0	0.0...	1
24	80.0	94.0	10...	76.0	76.0	94.0	94.0	72.0	72.0	81.0	82.0	68.0	76.0	83.0	91.0	74.0	71.0	79.0	87.0	70.0	71.0	79.0	79.0	67.0	75.0	83.0	85.0	71.0	71.0	75.0	85.0	67.0	71.0	79.0	77.0	67.0	7.0	0.0...	1
25	68.0	77.0	86.0	65.0	64.0	66.0	86.0	68.0	57.0	55.0	78.0	72.0	68.0	75.0	79.0	67.0	60.0	68.0	79.0	67.0	53.0	54.0	75.0	70.0	67.0	64.0	81.0	67.0	59.0	61.0	77.0	71.0	55.0	54.0	85.0	67.0	5.0	0.0...	1
26	64.0	66.0	86.0	68.0	57.0	55.0	78.0	72.0	53.0	49.0	71.0	65.0	60.0	68.0	79.0	67.0	53.0	54.0	75.0	70.0	53.0	54.0	71.0	63.0	59.0	61.0	77.0	71.0	55.0	54.0	85.0	67.0	55.0	51.0	74.0	67.0	5.0	0.0...	1
27	57.0	55.0	78.0	72.0	53.0	49.0	71.0	65.0	57.0	49.0	74.0	65.0	53.0	54.0	75.0	70.0	53.0	54.0	71.0	63.0	56.0	54.0	71.0	63.0	55.0	54.0	85.0	67.0	55.0	51.0	74.0	67.0	55.0	48.0	70.0	62.0	5.0	0.0...	1
28	84.0	10...	10...	81.0	84.0	99.0	10...	78.0	84.0	99.0	10...	81.0	84.0	10...	11...	87.0	84.0	99.0	10...	79.0	84.0	99.0	10...	79.0	95.0	12...	12...	10...	95.0	12...	12...	10...	82.0	10...	10...	85.0	3.0	0.0...	1
29	84.0	99.0	10...	81.0	76.0	99.0	10...	81.0	76.0	99.0	10...	85.0	84.0	99.0	10...	79.0	84.0	10...	10...	79.0	79.0	10...	10...	87.0	82.0	10...	10...	85.0	82.0	10...	10...	81.0	82.0	10...	10...	78.0	3.0	0.0...	1
30	80.0	10...	11...	88.0	80.0	11...	11...	88.0	80.0	10...	11...	85.0	79.0	10...	11...	87.0	79.0	10...	10...	83.0	79.0	10...	10...	79.0	82.0	10...	10...	85.0	82.0	10...	10...	85.0	82.0	10...	10...	85.0	3.0	0.0...	1
31	80.0	11...	11...	88.0	80.0	10...	11...	85.0	80.0	95.0	10...	78.0	79.0	10...	10...	83.0	79.0	10...	10...	79.0	79.0	95.0	10...	79.0	82.0	10...	10...	85.0	82.0	10...	10...	85.0	78.0	96.0	96.0	78.0	3.0	0.0...	1
32	80.0	91.0	10...	78.0	80.0	91.0	10...	74.0	80.0	95.0	10...	74.0	79.0	95.0	10...	75.0	75.0	95.0	10...	79.0	75.0	91.0	96.0	75.0	74.0	91.0	92.0	70.0	78.0	91.0	96.0	74.0	74.0	87.0	92.0	70.0	4.0	0.0...	1
33	80.0	95.0	10...	74.0	76.0	91.0	10...	74.0	76.0	95.0	10...	78.0	75.0	91.0	96.0	75.0	75.0	91.0	96.0	71.0	79.0	87.0	93.0	71.0	74.0	87.0	92.0	70.0	74.0	87.0	88.0	70.0	78.0	87.0	84.0	70.0	4.0	0.0...	1
34	76.0	91.0	10...	74.0	76.0	87.0	10...	74.0	76.0	87.0	91.0	74.0	79.0	87.0	93.0	67.0	75.0	87.0	96.0	71.0	75.0	91.0	96.0	71.0	74.0	87.0	88.0	66.0	74.0	87.0	92.0	70.0	78.0	87.0	88.0	66.0	4.0	0.0...	1
35	68.0	83.0	87.0	67.0	68.0	83.0	87.0	67.0	68.0	79.0	87.0	63.0	71.0	79.0	85.0	62.0	67.0	75.0	85.0	62.0	71.0	75.0	85.0	62.0	70.0	83.0	88.0	70.0	70.0	83.0	84.0	66.0	66.0	79.0	84.0	63.0	4.0	0.0...	1
36	84.0	10...	10...	85.0	84.0	10...	10...	85.0	88.0	10...	11...	88.0	79.0	99.0	10...	83.0	84.0	99.0	11...	87.0	84.0	99.0	10...	87.0	49.0	40.0	11...	11...	63.0	67.0	10...	85.0	82.0	96.0	10...	78.0	3.0	0.0...	1
37	88.0	10...	11...	92.0	88.0	10...	11...	92.0	88.0	11...	11...	88.0	84.0	10...	10...	83.0	88.0	10...	11...	87.0	88.0	10...	10...	87.0	86.0	10...	10...	85.0	90.0	10...	11...	85.0	86.0	10...	10...	85.0	3.0	0.0...	1
38	88.0	99.0	10...	85.0	88.0	10...	11...	88.0	88.0	11...	11...	92.0	88.0	10...	10...	87.0	84.0	10...	11...	87.0	88.0	11...	11...	92.0	90.0	10...	11...	89.0	90.0	10...	11...	85.0	90.0	10...	11...	85.0	3.0	0.0...	1
39	88.0	10...	11...	88.0	88.0	11...	11...	92.0	88.0	11...	12...	88.0	84.0	10...	11...	87.0	88.0	11...	11...	92.0	93.0	10...	10...	92.0	90.0	10...	11...	85.0	90.0	10...	11...	85.0	90.0	10...	11...	89.0	3.0	0.0...	1

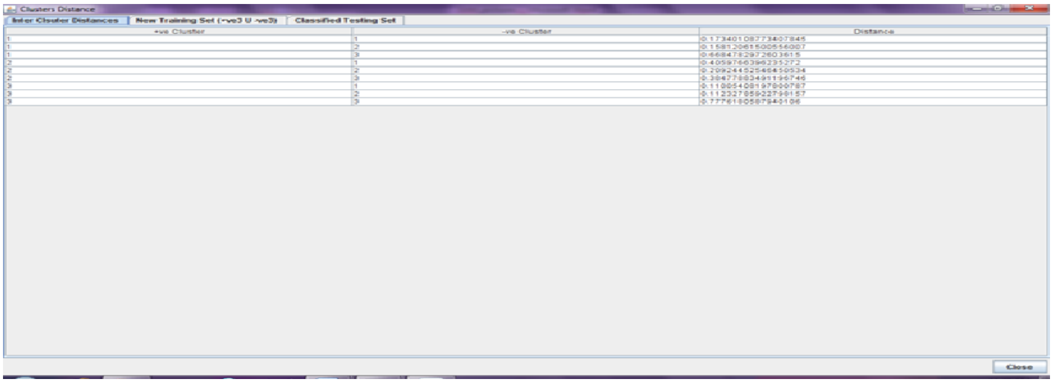
Fig-11. Different clusters with negative class



RID	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	f17	f18	f19
1	92.0	11...	12...	94.0	84.0	10...	10...	79.0	84.0	10...	10...	83.0	10...	12...	13...	10...	92.0	11...	11...
2	84.0	94.0	10...	79.0	80.0	94.0	98.0	76.0	80.0	10...	10...	79.0	84.0	99.0	10...	81.0	76.0	99.0	10...
3	80.0	94.0	98.0	76.0	80.0	10...	10...	79.0	76.0	10...	10...	79.0	76.0	99.0	10...	81.0	76.0	99.0	10...
4	76.0	89.0	98.0	76.0	76.0	94.0	98.0	78.0	76.0	98.0	10...	72.0	80.0	95.0	10...	74.0	76.0	91.0	10...
5	76.0	94.0	98.0	76.0	76.0	98.0	10...	72.0	76.0	94.0	90.0	76.0	76.0	91.0	10...	74.0	76.0	95.0	10...
6	75.0	98.0	10...	72.0	76.0	94.0	90.0	76.0	76.0	89.0	94.0	76.0	76.0	95.0	10...	78.0	76.0	91.0	10...
7	72.0	94.0	90.0	72.0	72.0	89.0	94.0	76.0	72.0	89.0	98.0	76.0	76.0	87.0	91.0	74.0	76.0	87.0	91.0
8	72.0	89.0	94.0	76.0	72.0	89.0	98.0	76.0	76.0	94.0	98.0	76.0	76.0	87.0	91.0	67.0	71.0	87.0	87.0
9	68.0	89.0	86.0	72.0	68.0	85.0	90.0	76.0	68.0	94.0	94.0	79.0	68.0	83.0	87.0	67.0	58.0	83.0	87.0
10	68.0	85.0	90.0	76.0	68.0	94.0	94.0	79.0	76.0	94.0	11...	79.0	68.0	83.0	87.0	67.0	71.0	83.0	87.0
11	84.0	98.0	11...	83.0	80.0	89.0	11...	87.0	88.0	10...	10...	87.0	88.0	10...	10...	85.0	84.0	99.0	10...
12	80.0	89.0	11...	87.0	88.0	10...	10...	87.0	92.0	11...	11...	91.0	84.0	99.0	10...	85.0	88.0	99.0	10...
13	92.0	11...	11...	94.0	92.0	11...	12...	91.0	84.0	10...	11...	87.0	88.0	11...	11...	92.0	88.0	11...	12...
14	84.0	98.0	10...	91.0	84.0	10...	11...	87.0	84.0	10...	11...	87.0	97.0	10...	11...	88.0	92.0	11...	11...
15	84.0	10...	11...	87.0	88.0	11...	11...	91.0	88.0	11...	12...	87.0	92.0	11...	11...	92.0	92.0	10...	11...
16	88.0	11...	11...	91.0	88.0	11...	12...	87.0	88.0	11...	11...	87.0	92.0	10...	11...	92.0	92.0	10...	11...
17	88.0	10...	10...	83.0	88.0	98.0	10...	83.0	84.0	10...	10...	83.0	84.0	99.0	10...	85.0	84.0	10...	10...
18	88.0	98.0	10...	83.0	84.0	10...	10...	83.0	88.0	11...	11...	87.0	84.0	10...	10...	81.0	88.0	99.0	10...
19	84.0	10...	10...	83.0	88.0	11...	11...	87.0	88.0	11...	10...	87.0	88.0	99.0	10...	85.0	84.0	10...	10...
20	88.0	11...	10...	87.0	88.0	10...	11...	87.0	84.0	10...	10...	87.0	88.0	95.0	10...	81.0	84.0	99.0	10...
21	84.0	10...	11...	87.0	88.0	10...	11...	87.0	88.0	10...	10...	87.0	88.0	10...	10...	85.0	88.0	10...	11...
22	88.0	10...	11...	87.0	88.0	10...	10...	87.0	84.0	10...	10...	87.0	88.0	10...	11...	85.0	88.0	99.0	10...
23	76.0	89.0	98.0	79.0	68.0	77.0	94.0	79.0	60.0	62.0	78.0	76.0	71.0	75.0	87.0	78.0	60.0	54.0	87.0
24	60.0	62.0	78.0	76.0	64.0	73.0	90.0	76.0	80.0	94.0	10...	83.0	56.0	61.0	87.0	78.0	71.0	79.0	10...
25	72.0	81.0	82.0	68.0	68.0	73.0	78.0	65.0	64.0	69.0	78.0	65.0	71.0	79.0	79.0	67.0	71.0	79.0	83.0
26	68.0	73.0	78.0	65.0	64.0	69.0	78.0	65.0	68.0	77.0	86.0	65.0	71.0	79.0	83.0	67.0	71.0	79.0	79.0
27	57.0	49.0	74.0	65.0	53.0	49.0	74.0	68.0	53.0	52.0	74.0	68.0	56.0	54.0	71.0	63.0	56.0	51.0	67.0
28	10...	12...	13...	10...	92.0	11...	11...	95.0	94.0	10...	10...	91.0	10...	12...	13...	10...	90.0	12...	12...
29	92.0	11...	11...	85.0	84.0	10...	10...	81.0	84.0	99.0	10...	78.0	88.0	12...	12...	10...	84.0	10...	11...
30	84.0	99.0	10...	78.0	84.0	99.0	10...	81.0	76.0	99.0	10...	81.0	84.0	99.0	10...	79.0	84.0	99.0	10...
31	76.0	99.0	10...	85.0	76.0	10...	11...	88.0	80.0	10...	11...	88.0	79.0	10...	10...	87.0	79.0	10...	10...
32	76.0	10...	11...	88.0	80.0	10...	11...	88.0	80.0	11...	11...	88.0	79.0	10...	10...	87.0	79.0	10...	11...
33	76.0	91.0	10...	74.0	76.0	95.0	10...	78.0	76.0	91.0	10...	74.0	75.0	91.0	95.0	71.0	79.0	87.0	93.0
34	76.0	95.0	10...	78.0	76.0	91.0	10...	74.0	76.0	87.0	10...	74.0	79.0	87.0	93.0	71.0	79.0	87.0	93.0
35	68.0	83.0	87.0	67.0	68.0	79.0	87.0	63.0	68.0	79.0	87.0	67.0	67.0	75.0	85.0	62.0	71.0	75.0	85.0
36	71.0	83.0	87.0	67.0	68.0	83.0	87.0	67.0	68.0	93.0	87.0	67.0	71.0	79.0	85.0	62.0	71.0	75.0	81.0
37	68.0	83.0	87.0	67.0	68.0	83.0	87.0	67.0	71.0	83.0	87.0	70.0	71.0	75.0	81.0	67.0	71.0	75.0	81.0
38	76.0	91.0	91.0	74.0	76.0	96.0	10...	81.0	84.0	10...	10...	85.0	67.0	75.0	96.0	79.0	75.0	83.0	96.0
39	76.0	95.0	10...	81.0	84.0	10...	10...	85.0	84.0	10...	10...	85.0	75.0	83.0	95.0	83.0	79.0	99.0	10...

RID	f20	f21	f22	f23	f24	f25	f26	f27	f28	f29	f30	f31	f32	f33	f34	f35	f36	f37	Alpha	y
1	85.0	84.0	10...	10...	81.0	10...	12...	13...	10...	88.0	12...	12...	10...	84.0	10...	11...	87.0	3.0	0.0...	-1
2	81.0	76.0	99.0	10...	85.0	84.0	99.0	10...	79.0	84.0	10...	10...	79.0	79.0	10...	10...	87.0	3.0	0.1...	-1
3	85.0	76.0	10...	11...	88.0	84.0	10...	10...	79.0	79.0	10...	10...	87.0	79.0	10...	10...	87.0	3.0	0.1...	-1
4	74.0	76.0	95.0	10...	78.0	75.0	91.0	96.0	75.0	75.0	91.0	96.0	71.0	79.0	87.0	93.0	71.0	4.0	0.5...	-1
5	73.0	76.0	91.0	10...	74.0	75.0	91.0	96.0	71.0	79.0	87.0	93.0	71.0	79.0	87.0	93.0	67.0	4.0	0.6...	-1
6	74.0	76.0	87.0	10...	74.0	79.0	87.0	93.0	71.0	79.0	87.0	93.0	67.0	75.0	87.0	96.0	71.0	4.0	0.6...	-1
7	67.0	71.0	87.0	87.0	70.0	75.0	91.0	96.0	71.0	75.0	87.0	93.0	67.0	71.0	87.0	89.0	67.0	4.0	0.8...	-1
8	70.0	71.0	83.0	87.0	67.0	75.0	87.0	93.0	67.0	71.0	87.0	89.0	67.0	71.0	79.0	81.0	62.0	4.0	0.8...	-1
9	67.0	71.0	83.0	87.0	70.0	71.0	75.0	81.0	67.0	71.0	75.0	81.0	62.0	67.0	75.0	85.0	71.0	4.0	0.9...	-1
10	70.0	76.0	91.0	91.0	74.0	71.0	75.0	81.0	62.0	67.0	75.0	85.0	71.0	67.0	75.0	96.0	79.0	4.0	0.8...	-1
11	85.0	88.0	99.0	10...	85.0	84.0	99.0	10...	83.0	88.0	10...	10...	87.0	88.0	10...	10...	87.0	3.0	0.0...	-1
12	85.0	88.0	10...	11...	88.0	88.0	10...	10...	87.0	88.0	10...	10...	87.0	84.0	10...	11...	87.0	3.0	0.0...	-1
13	88.0	92.0	11...	12...	92.0	88.0	11...	11...	92.0	93.0	10...	10...	92.0	93.0	11...	11...	92.0	3.0	0.0...	-1
14	92.0	92.0	11...	11...	92.0	93.0	10...	11...	87.0	93.0	10...	10...	87.0	88.0	10...	10...	92.0	3.0	0.0...	-1
15	92.0	92.0	10...	11...	88.0	88.0	10...	10...	92.0	88.0	10...	10...	87.0	88.0	10...	10...	87.0	3.0	0.0...	-1
16	88.0	88.0	10...	11...	88.0	88.0	10...	10...	87.0	88.0	10...	10...	87.0	88.0	10...	10...	87.0	3.0	0.0...	-1
17	81.0	88.0	99.0	10...	85.0	88.0	10...	10...	83.0	84.0	99.0	10...	87.0	79.0	99.0	10...	79.0	3.0	0.1...	-1
18	85.0	84.0	10...	10...	85.0	84.0	99.0	10...	87.0	79.0	99.0	10...	79.0	88.0	95.0	10...	79.0	3.0	0.1...	-1
19	85.0	88.0	95.0	10...	81.0	79.0	99.0	10...	79.0	88.0	95.0	10...	79.0	88.0	95.0	10...	80.0	3.0	0.1...	-1
20	85.0	88.0	10...	11...	85.0	88.0	95.0	10...	83.0	88.0	10...	10...	83.0	88.0	10...	10...	80.0	3.0	0.0...	-1
21	85.0	88.0	99.0	10...	85.0	84.0	10...	10...	83.0	84.0	99.0	10...	83.0	84.0	10...	10...	80.0	3.0	0.0...	-1
22	85.0	84.0	99.0	10...	85.0	84.0	99.0	10...	83.0	84.0	10...	10...	83.0	88.0	99.0	10...	79.0	3.0	0.0...	-1
23	74.0	56.0	61.0	87.0	78.0	79.0	91.0	10...	79.0	75.0	79.0	96.0	79.0	75.0	83.0	96.0	79.0	5.0	0.8...	-1
24	81.0	80.0	95.0	10...	85.0	75.0	83.0	96.0	79.0	84.0	99.0	10...	83.0	84.0	99.0	10...	80.0	5.0	0.7...	-1
25	67.0	71.0	79.0	79.0	63.0	71.0	79.0	77.0	67.0	71.0	75.0	81.0	67.0	67.0	72.0	81.0	67.0	7.0	0.9...	-1
26	63.0	68.0	75.0	79.0	67.0	71.0	75.0	81.0	67.0	67.0	72.0	81.0	67.0	67.0	64.0	81.0	67.0	7.0	0.9...	-1
27	63.0	53.0	51.0	67.0	67.0	55.0	48.0	70.0	62.0	51.0	48.0	70.0	67.0	51.0	48.0	70.0	67.0	5.0	0.8...	-1
28	10...	94.0	10...	11...	97.0	90.0	11...	11...	92.0	90.0	11...	12...	96.0	95.0	12...	12...	10...	3.0	0.0...	-1
29	87.0	84.0	99.0	10...	79.0	90.0	11...	12...	96.0	95.0	12...	12...	10...	95.0	12...	12...	10...	3.0	0.0...	-1
30	79.0	84.0	10...	10...	79.0	95.0	12...	12...	10...	82.0	10...	10...	85.0	82.0	10...	10...	81.0	3.0	0.0...	-1
31	87.0	79.0	10...	11...	87.0	82.0	10...	10...	78.0	78.0	10...	10...	81.0	82.0	10...	10...	85.0	3.0	0.0...	-1
32	87.0	79.0	10...	10...	83.0	78.0	10...	10...	81.0	82.0	10...	10...	85.0	82.0	10...	10...	85.0	3.0	0.0...	-1
33	71.0	79.0	87.0	93.0	67.0	74.0	87.0	88.0	70.0	78.0	87.0	84.0	70.0	74.0	87.0	88.0	66.0	4.0	0.7...	-1
34	67.0	75.0	87.0	96.0	71.0	78.0	87.0	84.0	70.0	74.0	87.0	88.0	66.0	74.0	87.0	92.0	70.0	4.0	0.7...	-1
35	62.0	67.0	79.0	81.0	62.0	70.0	83.0	84.0	66.0	66.0	79.0	84.0	63.0	66.0	79.0	88.0	66.0	4.0	0.9...	-1
36	67.0	71.0	75.0	81.0	62.0	70.0	79.0	88.0	66.0	66.0	71.0	88.0	70.0	59.0	60.0	96.0	81.0	4.0	0.9...	-1
37	62.0	67.0	75.0	85.0	71.0	68.0	71.0	88.0	70.0	59.0	60.0	96.0	81.0	56.0	49.0	10...	1.0...	4.0	0.9...	-1
38	83.0	79.0	99.0	10...	83.0	49.0	40.0	11...	11...	45.0	34.0	12...	12...	49.0	40.0	11...	11...	4.0	0.4...	-1
39	83.0	84.0	99.0	11...	87.0	45.0	34.0	12...	12...	49.0	40.0	11...	11...	63.0	67.0	10...	85.0	3.0	0.4...	-1

Fig-12. Clusters distance



+ve Cluster	-ve Cluster	Distance
1	1	0.173401082773407845
1	2	0.158112061500556007
1	3	0.6684782972603615
2	1	0.4059766396235272
2	2	0.209244525466850534
2	3	0.38477883491196746
3	1	0.11005408197800787
3	2	0.1123278592279157
3	3	0.7776180587940106

6. CONCLUSION

Here in this paper, machine intelligence is introduced. FSVM and Fuzzy C-Means algorithm along with Possibilistic K-means algorithm is discussed. WPC-COA, proposed algorithm, which helps in categorization of web pages is presented and experimented. KPCM algorithms avoid clusters coincidence and are less sensitive to outliers

REFERENCES

[1] J. C. Platt, *Sequential minimal optimization-A fast algorithm for training support vector machines, in advances in kernel methods-support vector learning*. Cambridge, MA: MIT Press, 1998.

[2] D. Gomez, J. Montero, and G. Biging, "Improvements to remote sensing using fuzzy classification, graphs and accuracy statistics," *Pure Appl. Geophys.*, vol. 165, pp. 1555-1575, 2008.

[3] E. O. Edgar, F. Robert, and G. Federico, *Support vector machines: Training and applications*: A.I. Memo No. 1602, C.B.C.L Paper No.144, March 1997, 2004.

[4] P. Scott and H. Lutz, "Comparing the results of support vector machines with traditional data mining algorithms, supported by Amica Life Insurance Company."

[5] W. Xiao-Hong, "Coll. of electr and inf eng, Jiangsu Univ, Zhenjiang. A possibilistic C- means clustering algorithm based on kernel methods," presented at the Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, Nov 2005.

Views and opinions expressed in this article are the views and opinions of the author(s), Review of Computer Engineering Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.