ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

Unsupervised Spectral Rare Class Ranking for Fraud Detection

P.Rajeshwari, D.Maheshwari

M.Phil Scholar, Dept. of Computer Science, Dr. Dr.N.G.P. Arts & Science College, Coimbatore, India

Assistant Professor/Head, Dept. of Computer Technology, Dr. Dr.N.G.P. Arts & Science, College, Coimbatore, India

ABSTRACT: A many data mining problems, obtaining labels is costly and time consuming, if not practically infeasible. Here implementing new unsupervised spectral ranking method for anomaly, the proposed SRA can generate anomaly ranking either with respect to the majority class or with respect to two main patterns. The spectral optimization in Spectral Ranking method for Anomaly (SRA) can be viewed as a relaxation of an unsupervised Support Vector Machine problem. In this research work concentrate on developing the rare class kernel model, the optimization algorithm, and extensive computational comparisons between AUC-based and error rate based rare class nonlinear kernel learning, as well as computational efficiency improvement of RankRC over RankSVM.

KEYWORDS: Spectral Ranking method for Anomaly, Support Vector Machine RankSVM, RankRC, nonlinear kernel learning.

I. INTRODUCTION

Data mining is defined as the process of extracting previously unknown, valid, and actionable information from large databases and then using the information to make crucial business decisions. Finding information hidden in data is as theoretically difficult as it is practically important.

RARE CLASS

Recently there are major changes and evolution has been done on classification of data. Class imbalance problem become greatest issue in data mining. Imbalance problem occur where one of the two classes having more sample than other classes. The most of algorithm are more focusing on classification of major sample while ignoring or misclassifying minority sample.

II. LITERATURE SURVEY

L.M.Taft et.al [1] applied SMOTE in imbalanced data that severe from high sparsity in addition to high class skew. The objective of the present study s to apply Synthetic Minority over Sampling Technique (SMOTE) as an enhanced sampling method in a sparse dataset to generate prediction models to identify ADE in women admitted for the labor and delivery based on patient risk factors and comorbidities.

Ming Gao et.al [2] proposed a technique that combined (SMOTE) and the particle swarm optimization (PSO) and radial basis function (RBF) classifier. They applied PSO algorithm to determine the structure and the parameters of RBF kernels. In this contribution, proposed a powerful and efficient algorithm for solving two class imbalanced problems, referred to as the SMOTE+ PSO-RBF, by combining the SMOTE and the PSO optimised RBF classifierThe results explained the competitive performance of SMOTE+PSO.

Aditya Tayal[4], et.al nonlinear kernel based classification methods expressed as a regularized loss minimization problem. We address the challenges associated with both rare class problems and large scale learning, by optimizing area under curve of the receiver of operator characteristic in the training process, instead of classification accuracy and using a rare class kernel representation to achieve an efficient time and space algorithm. We provide justifications for the rare class representation and experimentally illustrate the effectiveness of RankRC in test performance, computational complexity, and model robustness. Comparing the RankSVM while performing similarly that AUC

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

measure, In addition, RankRC is computationally significantly more efficient with respect to both time and space requirements.

Bee Wah Yap, Khatijahhusna Abd Rani et.al [6] Proposed lot of methods. This paper applied four methods: Oversampling, under sampling, Bagging and Boosting in handling imbalanced datasets. Document classification, loan default prediction, fraud detection or medical Classification which involve a binary response variable, the dataset are often highly imbalanced. For a binary response variable with two classes, when the event of interest is underrepresented, it is referred to the positive or minority class. Thus, the number of cases for the negative or majority class is very much higher than the minority cases. A simulation study should be carried out whereby data are generated and then the different approaches are compared so as to obtain a conclusive decision on the best strategy to handle imbalanced data.

Mahendra Sahare, et.al. [7] Proposed lot of technique to solving multivariate problem. Implementation of binary classifier in the form of liner classifier generate such a problem, the first approach relied on extending binary classification problems to handle the multiclass case directly. This included neural networks, decision trees, support vector machines, naive bayes, and k-nearest neighbours. The second approach decomposes the problem into several binary classification tasks. Several methods are used for this decomposition: one versus- all, all-versus-all, error-correcting output coding, and generalized coding. The third one relied on arranging the classes in a tree, usually a binary tree, and utilizing a number of binary classifiers at the nodes of the tree till a leaf node is reached. In future it minimized the problem of feature reduction problem and error correcting code for binary classifier.

Francisco Fernández-Navarro et.al [9] proposed two oversampling methods; static SMOTE radial basis function method and a dynamic SMOTE radial basis function procedure incorporated into a memetic algorithm that optimizes radial basis functions neural networks. The experiments showed the highest accuracy and sensitivity level of the dynamic oversampling method comparing to other neural networks methods. A dynamic over-sampling procedure based on sensitivity for multi-class problems.

Gilles Cohen et.al [10] proposed a resampling approach using both oversampling and under-sampling with synthetic instances. Also, they introduced class-dependent regularization parameters for tuning SVM and obtained asymmetrical soft margin (larger margin on the side of the smaller class).The main objective to retrospective analysis of a prevalence survey of NIs in the Geneva University Hospital.

Breiman et.al [11] proposed concept of bootstrap aggregating to construct ensembles. That is, a new data-set is formed to train each classifier by randomly drawing instances from the original data-set. Hence, diversity is obtained with the re-sampling procedure by the usage of different data subsets. Finally, when an unknown instance is presented to each individual classifier, a majority or weighted vote is used to infer the class.

Hung-Yi Lin et.al [13] proposed multivariate statistical analyses. Multivariate statistical analyses have two advantages. First, they can explore the relationships between variables and find the most characterizing features of the observed data. Second, they can solve problems which are stalled by high dimensionality. Our learning model advances three new distinguishing characteristics including evaluation method, feature selection, and feature extraction. Hence, the main contributions of this paper are threefold. First, the enhanced relevance analysis is proposed for feature evaluation process. Second, the synergistic classification effect is enhanced by our heuristic feature selection algorithm. Finally, the generation of coarse-grained classifier composed of lowly correlated relevant features is successfully realized.

III. UNSUPERVISED SPECTRAL RARE CLASS RANKING

A. *SPECTRAL CLUSTERING:*

Spectral clustering has become a widely used clustering technique, often outperforming traditional clustering techniques such as k-means and hierarchical clustering. Before proposing our ranking method, we first briefly review the spectral clustering technique. The main objective of clustering is to partition data into groups so that similarity between different groups is minimized.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

B. ***SPECTRAL ANALYSIS AND CLUSTERING:***

Spectral clustering is a more recent, popular, and successful clustering technique, which often outperforms traditional clustering techniques such as k-means and hierarchical clustering the main objective of clustering, is to partition data into groups so that similarity between different groups is minimized. Hence similarity based clustering can be modelled as a graph cut problem. Let each data instance be a vertex and each pair of vertices be connected with an edge with a weight equal to the similarity of the pair. We can then represent the data and its similarity using an undirected graph G = (V E), with vertices V = fv1; v2, representing corresponding to data instances and an adjacency matrix summarizing similarities, where W_{ij} is the similarity between $v_i$andv_j. Let d be the degree vector of each vertex with d_j =$\sum jW_{ij}$ and D be the diagonal matrix with d on the diagonal. From the degree matrix D and the weighted adjacency matrix W, a Laplacian matrix, which is fundamental in spectral clustering computation, can be introduced. There are different variations in the definition of Laplacian. The relevant definition to our discussion in this paper is the symmetric normalized Laplacian.

$$L = I - D^{-1/2} W D^{-1/2} \quad \text{..... } \mathbf{1}$$

The approach is divided into two phases as follows: 1) Clustering phase: In this phase, clustering is applied to all instances of each class to detect class subtypes. This phase helps in finding the hidden patterns within each class and discovering more specific categories. Two clustering techniques are employed and compared namely K-means and hierarchical clustering. After that, instances belong to each cluster are relabelled with a new class. In this work, we vary of clusters from 2 to 5.

Algorithm 1

Classification algorithm

Step 1: Class Decomposition 1: for i=1 to |Y | do

Step 2: datai = instances of class i

Step 3: clusters = cluster(datai, k)

Step 4: for j=1 to k do

Step 5: dataj = instances of cluster j

Step 6: relabel(dataj)

Step 7: end for

Step 8: end for the classification and clustering.

In this algorithm |Y | as represents total number of classes. Then Classification phase: In this phase, classification is applied on data objects based on the new classes produced in the clustering phase.

K-means aims to divide the instances into k clusters; where each instance belongs to the cluster with the nearest mean. It aims to minimize the within-cluster sum of squares objective's-means is described by Algorithm.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

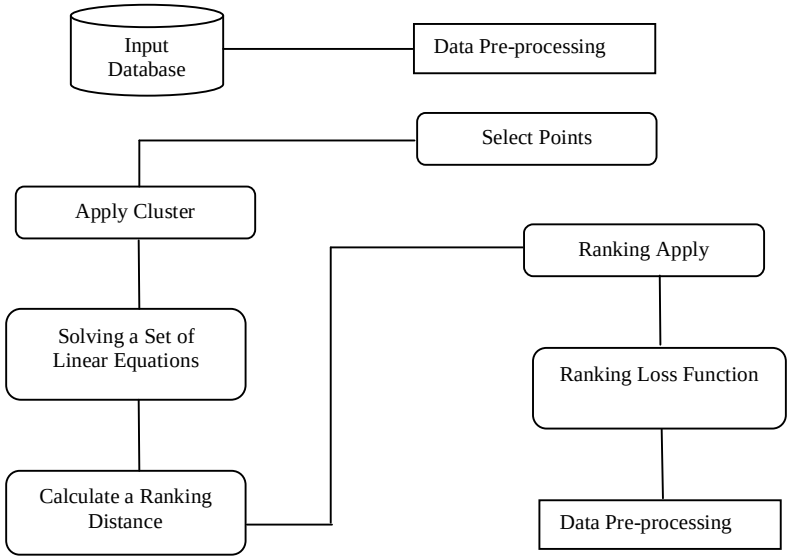


FIGURE 4.1 SPECTRAL RANKING METHODS FOR ANOMALY PROCESS

Spectral Clustering With Eigenvector Relevance Learning

Let us first formally define the spectral clustering problem. Given a set of *N* data points/input patterns represented using feature vectors

$$D = \{f_1. . . , f_n, . . . , f f_{1N}\} \text{.....} \mathbf{2}$$

We aim to discover the natural grouping of the input data. The optimal number of groups/clusters K_0 is automatically determined to best describe the underlying distribution of the dataset. We have $K_0 = K_{true}$ if it is estimated correctly. Note that different feature vectors can be of different dimensionalities. An $N \times N$ affinity matrix A = $\{A_{ij}\}$ can be formed whose element A_{ij}measures the affinity/similarity between the *i*th and *j*th feature vectors. Note that A needs to be symmetric, i.e. $A_{ij}j = A_{ji}$. The eigenvectors of A can be employed directly for clustering. However t it is more desirable to perform clustering based on the eigenvectors of the normalised affinity matrix$\bar{A}$, defined as

$$\bar{A}=L^{-\frac{1}{2}} A L^{-\frac{1}{2}} \text{ .....} \mathbf{3}$$

Where L is an $N \times N$ diagonal matrix with $=L_{ii} \sum j A_{ij}$We assume that the number of clusters is between 1 and K_m,a number considered to be sufficiently larger than K_0. The training data set is then represented in an eigenspace using the *Km* largest eigenvectors of. $\bar{A}$ Denoted as

$$D_e = \{X_1, . . . , X_n, . . . , \{X_N\}, \text{.....} \mathbf{4}$$

With the *n*th feature vector f*n* being represented as a *Km* dimensional vector $X_n = [e_{1n}, . . . , e_{kn}, . . . , e_{1Km} n,]$, where . , $e_{kn}$is the *n*th element of the K_{th} largest eigenvector e_k. Note that now each feature vector in the new data set is of the same dimensionality *Km*. the task of spectral clustering now is to determine the number of clusters and then group the data into different clusters using the new data representation in the Eigen space.

Algorithm 2

Spectral Ranking for Abnormality (SRA)

Step 1: input: W: an n-by-n similarity matrix W

Step 2: X: ratio upper bound on anomaly

Step 3: Output$f^* \in R^n$: a ranking vector with a larger value representing more abnormal

Step 4: mFLAG : a flag indicating the type of ranking reference
begin

Step 5: Form Laplacian L =I-$D^{-1/2} W D^{-1/2}$

Step 6: COMPUTE $z^*=D^{\frac{1}{2}} g_1^*$ and g_1^*(the first non- principal eigenvector for L);

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

Step 7: Let $C_+=\{i: z_i^* \geq 0\}$ and $C\_=\{i: z_i^* < 0\}$;
Step 8: if $\min\left\{\frac{|C_+|}{n}, \frac{|C\_|}{n}\right\} \geq X$ then
Step 9: mFLAG= 1,f^*max($|z^*|$)- $|z^*|$,% ranking w.r.t. multiple pattern;
else if $|C_+| > |C\_|$ then
Step 10: mFLAG= 0, $f^* = -z^*$, %ranking w.r.t. Single major pattern;
else
Step 11: mFLAG= 0,$f^* = z^*$, %ranking w.r.t. Single major pattern;
end
end

If one of C+ and C- is sufficiently small in size relative to the other In proposed SRA provides anomaly ranking score f* with respect to one majority class (rare class ranking). Otherwise SRA yields an anomaly ranking with respect to the two $(\pm)$ major classes. Specifically, let $C_+=\{i, (g_1)_i^* \geq 0\}$and $C\_=\{i, (g_1^*)i < 0\}$ denote data instance index sets corresponding to non-negative and negative value in (g_1^*) respectively. In addition, assume that an a priori upper bound for the anomaly ratio X is given. if $\min\left\{\frac{|C_+|}{n}, \frac{|C\_|}{n}\right\} \geq X$ then then neither the set C_+nor $C_$is considered as an anomaly class and SRA outputs ranking with respect to multiple patterns. Otherwise, SRA outputs anomaly ranking with respect to a single majority class. If the 1st non-principal eigenvector is relatively balanced, $|C_+| \approx |C\_|$ SRA output anomaly ranking with respect to multiple patterns.

IV. IMPLEMENTATION AND DATA ANALYSIS

Data Set: Confusion matrix defines four possible scenarios when classifying class "C" Predicted Class "C" Predicted Class "NC" Actual class "C" True Positives (TP) False Negatives (FN) Actual class "NC" False Positives (FP) True Negatives (TN) Predicted Class "C" Predicted Class "NC" Actual class "C" True Positives (TP) False Negatives (FN) Actual class "NC" False Positives (FP) True Negatives (TN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F-value = Re call Precision ( ) Re call Precision

Take the difference between a feature vector (minority class sample) and one of its k nearest neighbours (minority class samples). Multiply this difference by a random number between 0 and 1. Add this difference to the feature value of the original feature vector, thus creating a new feature vector.

ŠDetection rate (Recall) - ratio between the number of correctly detected rare events and the total number of rare events ŠFalse alarm (false positive) rate – ratio between the number of data records from majority class that are misclassified as rare events and the total number of data records from majority class ŠROC Curve is a trade-off between detection rate and false alarm ra.

Model	Accuracy	Detection rate	Recall
SVM	0.90	0	0.91
Navie Bayes	0.85	0	0.91
K-Means	0.89	0	0.90
Rare Class	0.92	0.33	0.95
Proposed Framework	0.93	0.35	0.96

TABLE 5.1 COMPARISONS BETWEEN EXISTING AND PROPOSED ALGORITHM

Description of the German credit dataset

- 1. Title: German Credit data
- 2. Number of Instances: 1000
- 3. Number of Attributes German: 20 (7 numerical, 13 categorical)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

- Number of Attributes german.numer: 24 (24 numerical).

V. RESULT ANALYSIS

Data Pre-Processing

Data pre-processing involves analyzing incomplete, noisy and inconsistence data are commonplace properties of large real-world database.

EData Selection Form File

In this figure 5.1 to represent the select the original data from input file. Here that data file name is german data. Then preprocssing the german data. In this german data will be common for all credit card data set. The data contains data on 20 variables and the classification whether an applicant is considered a Good or a Bad credit risk for 1000 loan applicants."Let's look at the variables in the data set"

```
 [1] "account.status"   "months"           "credit.history"
 [4] "purpose"          "credit.amount"    "savings"
 [7] "employment"       "installment.rate" "personal.status"
[10] "guarantors"       "residence"        "property"
[13] "age"              "other.installments" "housing"
[16] "credit.cards"     "job"              "dependents"
[19] "phone"            "foreign.worker"   "credit.rating"
```

There are 1000 of record will be pre-processed. According to that figure A11, A12, A14.., will be represents the numerical attribute. After that 6,48..., represent to the credit card features like as account type, location, salary detail and all other details.In this figure5.1 shows identify the some data's are unnecessary. It performed to remove the unnecessary feature extraction method.

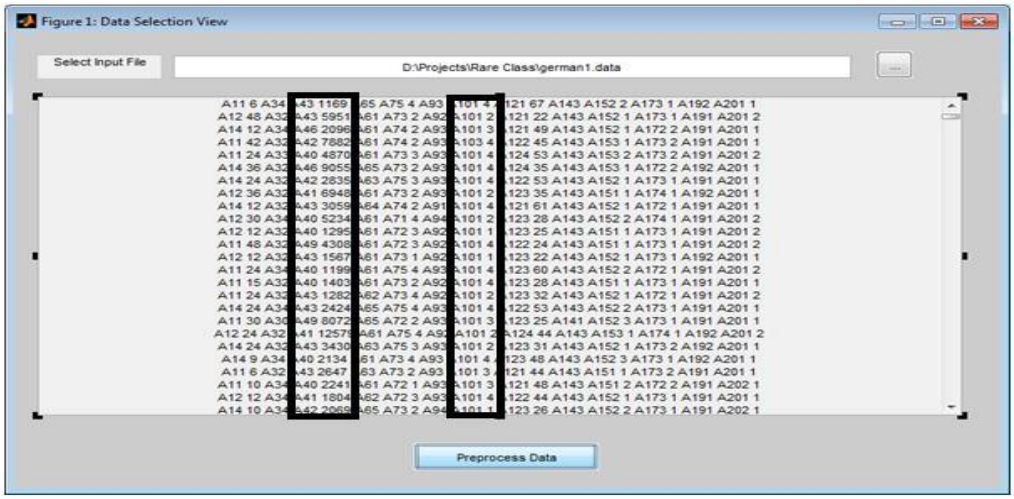


Figure 5.1 Pre-Processing

Feature Selection

After complete this pre-processing data will collecting many features, it will take more time. So, select the particular feature. Feature extraction technique is used to extract the subset of new features from the original feature set by means of some functional mapping by keeping as much information in the data as possible. Eigen vector formula used the feature extraction process.

Original Data

In this figure 5.2 represent original data. It has related credit card feature. It will be shown the salary in account details, and the number of month salary detail. Mostly credit card fraud depaed to the location fraud type and timing detail.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

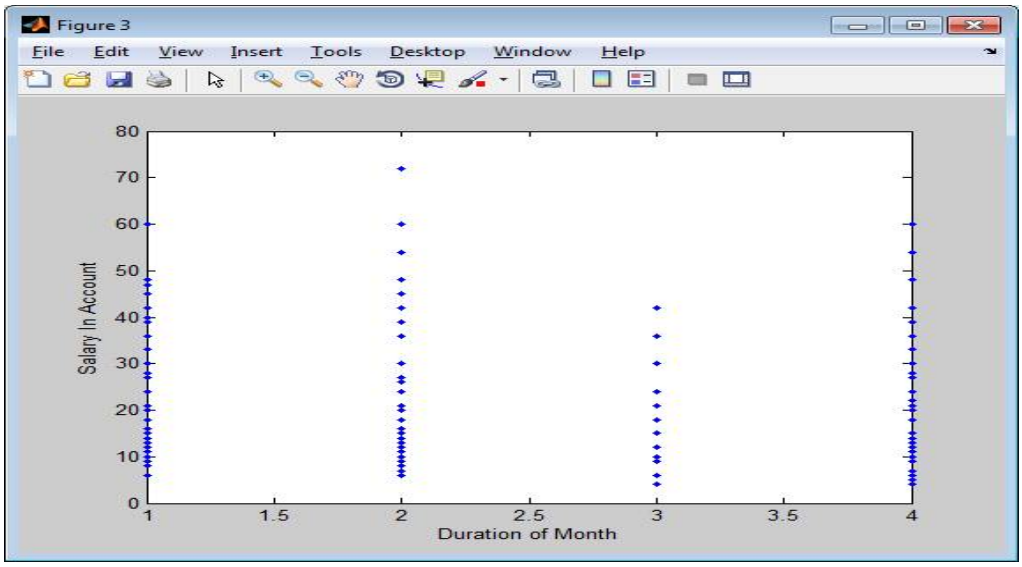


Figure 5.2 Original Data

Spectral Clustering Using Laplacian Method

To apply the clustering involves Laplacian method. In this method used to identify the features.

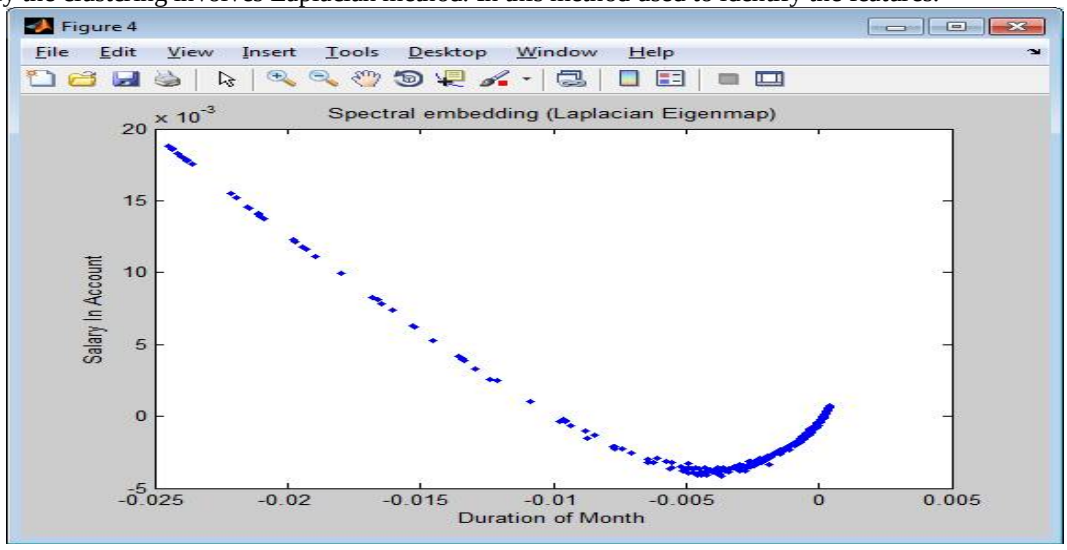


FIGURE 5.3 SPECTRAL CLUSTERING USING LAPLACIAN METHOD

Ranking Distance

In this figure 5.4 shown the ranking distance. Rank SVM methods mention the low level rare class score. And RankSVM with cluster identify the high level rare class score. After calculating ranking distance then apply the ranking.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

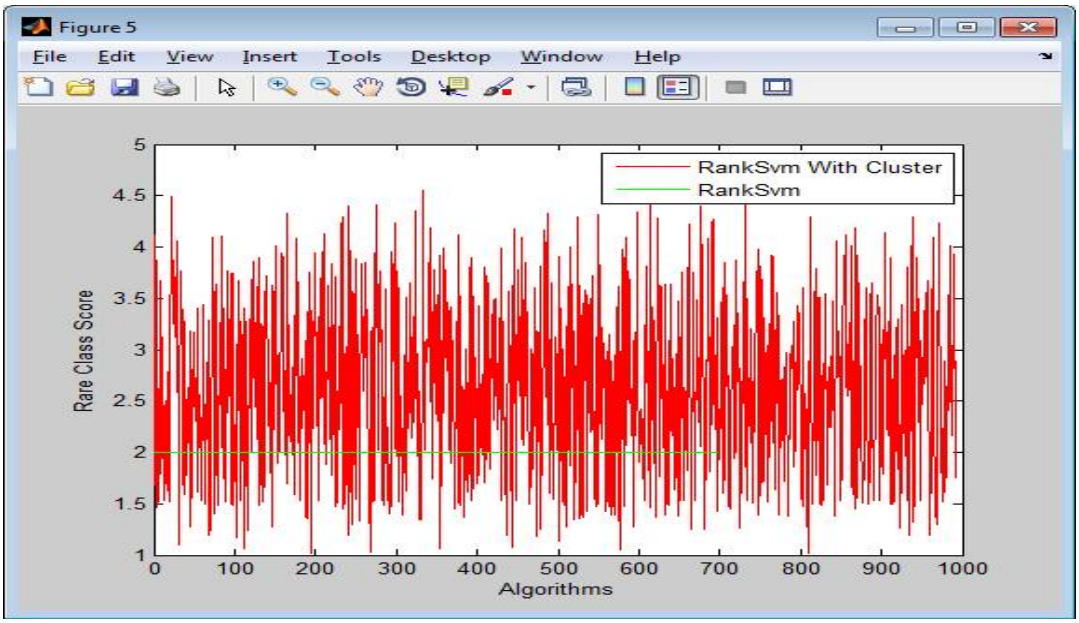
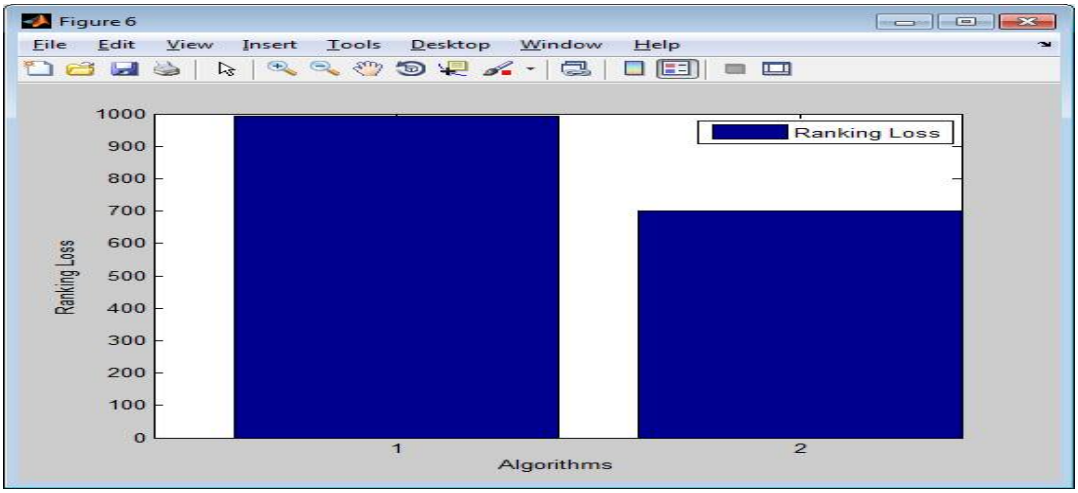


FIGURE 5.4 RANKING DISTANCE

RANKING LOSS FUNCTION

In this figure 5.5 shown the ranking loss function using spectral clustering algorithm and rank SVM algorithm.The rank SVM method make high level ranking loss as well as spectral clustering method show the the low level ranking loss.



5.5 RANKING LOSS FUNCTION

VI. CONCLUSION AND FUTURE WORK

In this research work has been proposed spectral ranking method for anomaly detection. It observed that the spectral optimization problem can be interpreted as an approximation to an unsupervised support vector machine and a non-principal eigenvector can be used to derive a ranking vector directly. It allowed a choice of the reference in the assessment of anomaly ranking. If the minority class does not have a sufficiently large count percentage, one can

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

choose to assess anomaly likelihood with respect to a single majority class and ranking is generated suitably with this view. This method uses unique oversampling technique to almost balance dataset such that to minimize the "uniform effect "in the clustering process combined with ranking loss function Based on this information in an eigenvector, a data instance is more likely to be an anomaly if its magnitude is smaller, when both positive and negative classes cannot be ruled as abnormal based on instance. It considers the challenging credit fraud detection problem based on a real claim data set. Since obtaining such labels are time consuming, costly and error prone in real applications, In this model the problem unsupervised learning and ignore labels when generating ranking using SRA, even though fraud labels are given for this particular data set.

Future Enhancement

1) In future research to identify new classes quickly and active learning to train classifiers with minimal supervision. These goals occur together in practice and are intrinsically related because examples of each class are required to train a classifier.
2) Active learning algorithm to optimise the data mining for rare classes in new domains makes inefficient use of human supervision.
3) Developing active learning algorithms to optimise both rare class discovery and classification simultaneously is challenging because discovery and classification have conflicting requirements in query criteria.
4) A unified active learning model to jointly discover new categories and learn to classify them by adapting query criteria online; and a classifier combination algorithm that switches generative and discriminative classifiers as learning progresses

REFERENCES

1. L.M.Taft "Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery", science direct, Vol: 42,pp: 356 –364, 2009.
2. Ming Gao" A combined SMOTE and PSO based RBF classifier for two class imbalanced problems" Elsevier, Vol: 74, pp: 3456-3466, 2011.
3. Yang Yong "The Research of Imbalanced Data Set of Sample Sampling Method Based on K-Means Cluster and Genetic Algorithm" ,Elsevier, Vol: 17,pp:164 –170,2012.
4. Aditya Tayal, Thomas F. Coleman, and Yuying Li" RankRC: Large-Scale Nonlinear Rare Class Ranking" IEEE transactions on knowledge
5. Yukio Ohsawa, Hiroyuki Kido, Teruaki Hayashi, and Chang Liu" Data Jackets for Synthesizing Values in the Market of Data" Procedia Computer Science, Vol: 22, pp: 709 -716, 2013.
6. Clifton Phua, Vincent Lee1, Kate Smith1 and Ross Gayler" A Comprehensive Survey of Data Mining-based Fraud Detection Research".
7. Mahendra Sahare and Hitesh Gupta," A Review of Multi-Class Classification for Imbalanced Data", International Journal of Advanced Computer Research, Vol:2,Issue:3 pp:163-168,2012.
8. Anuj Sharma and Prabin Kumar Panigrahi "A Review of Financial Accounting Fraud Detection based on Data Mining Techniques" International Joue in erasrnal of Computer Applications, Vol: 39, Issue: 2, pp: 37 – 47, 2012.
9. Francisco Fernández-Navarro"A dynamic over-sampling procedure based on sensitivity for multi-class problems" Elsevier, Vol: 44, Issue: 8, pp: 1821–1833, 2011.
10. Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel and J. Christopher Westland" Data mining for Credit Card Fraud: A comparative study" Decision Support Systems, Vol: 50,pp: 602 – 613, 2011.
11. L. Breiman, "Bagging predictors" Machine Learning, vol: 4, pp: 123–140, 1996.
12. B. Raskutti and A. Kowalczyk, "Extreme re-balancing for SVMS: A case study," SIGKDD Explor. Newslett., vol. 6, pp. 60–69,2004.
13. G. Wu and E. Y. Chang, "Class-boundary alignment for imbalanceddataset learning," in Proc. Int. Conf. Mach. Learn.,pp. 49–56, 2003.
14. G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data,"SIGKDD Explor. Newslett., vol: 6, pp. 20–29, 2004
15. Gilles Cohen "Learning from imbalanced data in surveillance of nosocomial infection" Elsevier, vol: 37, pp: 7-18, 2006.