Implementation of Outlier Detection Techniques for Imperfect Data Labels

Ms.Priyanka Meshram
Dept. of Computer science & Engineering,
G.H.Raisoni Institude of Engg. & Technology for Women,
Nagpur, India

Prof. Sapna Khapre
(Assistant Prof.)
Dept. of Computer Science & Engineering,
G.H.Raisonil Institute of Engg. &Technology For Women,
Nagpur , India

Prof.Priyanka Fulare
(Assistant Prof.)
Dept. of Computer science & Engineering,
G.H.Raisoni Institute Of Engg. & Technology For Women,
Nagpur,India

Abstract— A dataset may contain objects that do not comply with the general behaviour or model of data. These data objects are outlier. Many data mining methods discard outliers as noise or exceptions.However in some applications the rare events can be more instersting than the more regularly occurring ones.The analysis of outlier data is refered as outlier analysis or analysis mining.With datalables are imperfect i.,e inconsistent data as our database is collection of huge data, so that we are arranging the data in proper formats. Our objective would be to actively select instances with higher probabilities to be informative in determining feature relevance so as to improve the performance of feature selection without increasing the number of sampled instances. In this project, we would be applying a combination of canopy clustering,clustering weighted modeling and data cube clustering algorithms to get a better output for text classification with respect to the methods available.We will perform a comparative study on a variety of feature selection methods for data clustering, with other algorithms.Finally, we will evaluate the performance of hybrid feature selection method based on clustering.

Keywords*- Outlier Detection, Data of uncertainty, Feature Selection*

*****

I. INTRODUCTION

As the dataset contain the objects ,the task of outlier detection is to identify data object that are deviated from other data.The purpose of feature selection is to determine which features are the most relevant to the current classification task. In data classification, features are typically words from a document. Choosing an appropriate feature selection method for data classifcation can be vital because of the large number of features usually present in documents.Priviously support vector data iteration method used for outlier detection which is capable of detecting oulier in various application domains.Outlier analysis may uncover frandulent usage of cards by detecting purchase of unusually large amounts for a given account number in comparision to regular charges incurred by the same account.Outlier value may as to be detected with respect to the locations and types of purchase, or the purchase frequency.In this project hybrid algorithm will using which is consisting of three types of algorithms namely as canopy clustering,clustering weighted and data cube clustering,which will provide maximum accuracy and less searching time.Our main objectives are

- To find regression of existing data for maximum accuracy
- Design hybrid algorithm which is more efficient than existing algorithm
- Testing the algorithm on different datasets,gives more accurate result analysis

The proposed scheme overcomes the drawbacks of existing scheme such as inefficiency and inaccuracy. It provides less search time and high retrieval accuracy.

II. LITERATURE REVIEW

In this section, a review of previous work on outlier detection is presented. An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. While outlier detection aims to identify errors and remove their contaminating effect on the data set and as such to purify the data for processing. The different methods for outlier detection are given below,

- Bo Liu, Yanshan Xiao, Philip S. Yu, Zhifeng Hao, and Longbing Cao [1] proposes a novel outlier detection approach to address data with imperfect labels and incorporate limited abnormal examples into learning. To deal with data with imperfect labels, introduce likelihood values for each input data which denote the degree of membership of an example toward the normal and abnormal classes respectively.
- Chang-Dong Wang, Jian-Huang Lai, Dong Huang, and Wei-Shi Zheng [2] proposes a novel data stream clustering algorithm, termed SVStream, which is based on support vector domain description and support vector clustering.
- Y. J. Lee, Y. R. Yeh, and Y. C. F. Wang [3] proposes an online oversampling principal component analysis (osPCA) algorithm to address this problem, and aim at detecting the presence of outliers from a large amount of data via an online updating technique.
- Xiaoli Li, Chris P. Bowers, and Thorsten Schnier [4] proposes an intelligent data analysis method for modeling and prediction of daily electricity consumption in buildings.
- C.C.Aggarwal and P.S.Yu [5] provide a survey of uncertain data mining and management applications. Explore the various models utilized for uncertain data representation. In the field of uncertain data management,

examine traditional database management methods such as join processing, query processing, selectivity estimation, OLAP queries, and indexing.

2.1 EXISTING SYSTEMS

The detail review of efficient approach for outlier detection has been given by Bo Liu, Yanshan Xiao,Philip S. Yu, Zhifeng Hao, and Longbing Cao. There are different approaches for outlier detection.

2.1.1 Support vector data iteration

The support vector data iteration (SVDI) has been proposed for one-class classification learning. Given a set of target data {**x***i*}, *i* = 1, . . , *l*, where **x***i* ∈ *Rm*, the basic idea of SVDI is to find a minimum hyper-sphere that contains most of target data in the feature space, as illustrated in fig.1

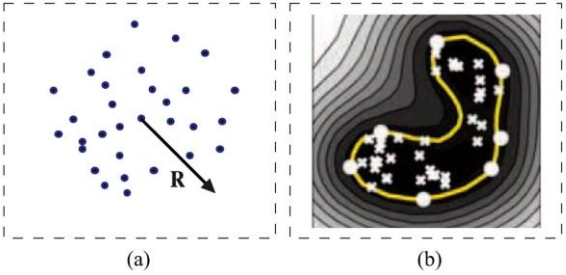


Fig. 1.(a) Illustration of SVDI hyper-sphere in feature space.(b) Illustration of SVDI decision boundary in input space.

$$\text{Min } F(R,o,\ i) = R^2 +\ C \qquad ,$$

$$\text{s.t. } \| \phi(xi) - o \|^2 <= R^2 + \qquad (1)$$

where *φ(.)* is a mapping function which maps the input data from input space into a feature space, and *φ(*x*i)* is the image of **x***i* in the feature space, *ξi* are slack variables to allow some data points to lie outside the sphere, and *C* > 0 controls the tradeoff between the volume of the sphere and the number of errors. is the penalty for misclassified Samples.

By introducing Lagrange multipliers *αi*, the optimization problem (1) is transformed into:

$$\max \sum_{i=1}^{l} \alpha i K(xi, xi) - \sum_{i=1}^{l} \sum_{k=1}^{l} \alpha i \alpha\, k K(xi, xk)$$

$$\text{s.t. } 0 <= \alpha i <= C, \qquad (2)$$

in which kernel function *K(, .,)* is utilized to calculate the inner pairwise product of two vector *φ(***x***i)* and *φ(***x***j)*, that is *K(***x***i*, **x***j)* = *φ(***x***i)* ·*φ(***x***j)*. The samples with *αi* > 0 are support vectors (SVs). For a test point **x**, it is classified as normal data when this distance is less than or equal to the radius *R*. Otherwise, it is flagged as an outlier.

2.2.2 Kernel k-Means clustering-based method

We adopt the kernel *k*-means clustering algorithm to generate likelihood values for each input data. In kernel-based method, a nonlinear mapping function *φ(.)* maps the input samples into a feature space. Kernel *k*-means clustering minimizes the following objective function

$$J = \sum_{i=1}^{k} \sum_{j=1}^{l+n} | ^2, \qquad (3)$$

where *k* is the number of clusters and **v***i* is the cluster center of the *ith* cluster. By solving this optimization problem, *k*-means clustering returns a set of local clusters, in which data samples belonging to a same cluster are more similar to each other. Intuitively, for a data sample, if most of data samples in the same cluster are normal, it would have a high probability of being normal, and if there is an outlying point that does not belong to any cluster, it would have a high probability of being an outlier. Therefore, we calculate the likelihood values for single likelihood model and bi-likelihood model as follows. For a given cluster *j*, assume there exist l^p_j normal examples and l^n_j negative examples.

2.2.3Kernel LOF-based method

To cope with datasets with varying densities, we propose a local density-based method to compute likelihood values for each input data. Inspired by the LOF algorithm , the basic idea is to examine the relative distance of a point to its local neighbors in feature space. More specifically, we extend the original LOF into the kernel space by using kernel function and generate the likelihood values in the kernel space instead of the input space.The key idea is that every object in a data set is an outlier to some extent and this extent is measured using the Local Outlier Factor (LOF).

III. TECHNIQUES ELABORATED FOR OUTLIER DETECTION

In this paper, inspired by support vector iteration iteration (SVDI) and support vector clustering , we propose a novel data stream clustering algorithm termed Support Vector-based Stream clustering (SVStream). According to the SVDI theory,support vectors can obtain flexible and accurate data descriptions by mapping the data into a kernel space. These SVs are used to construct cluster boundaries of arbitrary shape in SVC.

- SVStream accurately discovers clusters of arbitrary shape by constructing cluster boundaries using SVs.
- It can adapt to dramatic and gradual changes in an evolving stream by dynamically maintaining multiple spheres.
- It can discover overlapping clusters by allowing for bounded support vectors (BSVs). It is also effective in detecting and removing outliers (noise) via the BSV decaying mechanism.
- It is very efficient and requires very small memory space due to the compact representation of the summary information by multiple spheres.

In this section, we provide a detailed description about our proposed approaches to outlier detection.Given a datasets of USA hospital as examples, the objective is to classify the hospital data according to the various cities comes under USA. However, subject to find out regression calculation, an normal example may behave like an outlier, even though the example itself may not be an outlier.Based on outlier factor of cluster,a clustering-based outlier detection method, named CBOD is presented .The method consist of two stages,first stage cluster the datasets by one pass clustering algorithm and second stage determine outlier cluster by outlier factor.Time complexity is same as the size of datasets. The main objective of this proposed scheme is to overcome the drawbacks of existing scheme such as inefficiency , inaccuracy and noise ratio. Based on this, our proposed approaches work in Following steps as follows:

3.3.1 Data Clustering

Data clustering means grouping of similar types of data.Here support vector data description has been demonstrated to be capable of detecting outlier in various application domains.In this we are having large datasets of USA hospital.then grouping according on the basis of attributesor keyword ,based on this result found only that type of data from whole datasets of USA hospital.Data clustering is performed based on K-means algorithm.

3.2 Data Indexing

The basic need of indexing is to remove ambiguity in data,so that by using indexing provide id of the resulting data.BR tree method is used for indexing,in this method perform 1000 times clustering on resulting data.Here after performing the clustering,indexing is performed only on resulting clustering.By performing indexing provide id,this provide primary key which is used for further calculations and analysis on datasets.

3.3 Data Sequencing

After providing the indexing ,we are arranging the data according to the descending order.Data sequencing is used for segmentation.

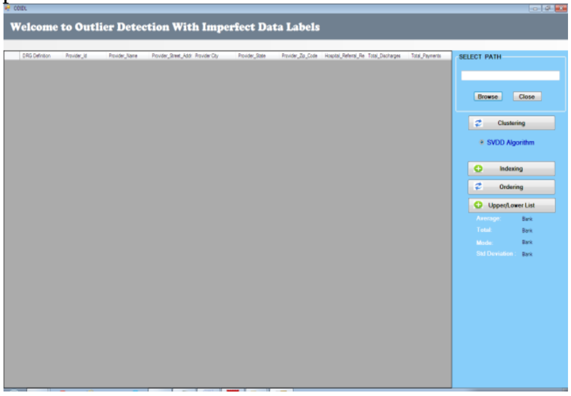
3.4 Data segmentation

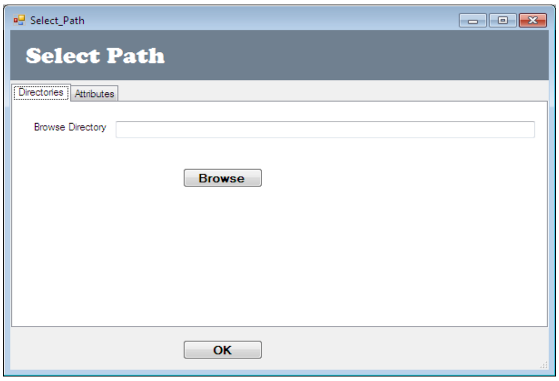
Data segmentation is used for data analysis, devide the result into two list first one is upper list and second one is lower list.upper list having decreasing order of data and lower list having increasing order of data.then after this we are calculating regression ratio. Regression calculation having following factors:Man value is used for average no. of total payment paid by patient,mode calculate no.of frequency,std deviation is summation of square of difference between each value and mean value to the total no. of value,provide how much it is differ from average calculating value.

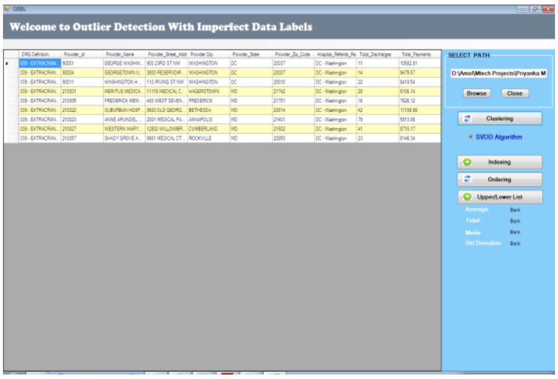
- Mean $= \sum \frac{x}{n}$
- Mode= L+(f1-f0/2f1-f0-f2)
- Std .deviation $=\sqrt{\sum(x - x')^2}/ n$

IV. RESULT EVALUATION

In this section, we conduct extensive result to investigate the performance of our proposed approach. For all reported results, the test platform is a Dual 2.4GHz Intel Core2 T9600 laptop with 4GB RAM

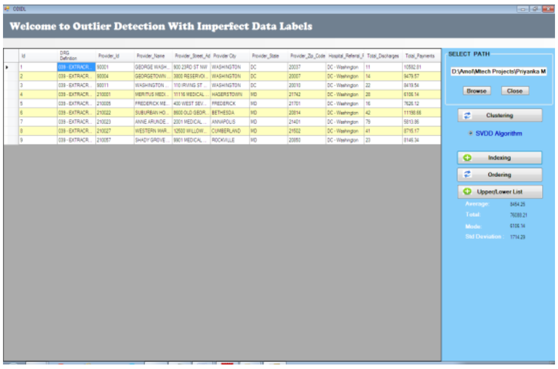






D:\Amol\Mtech Projects\Priyanka M

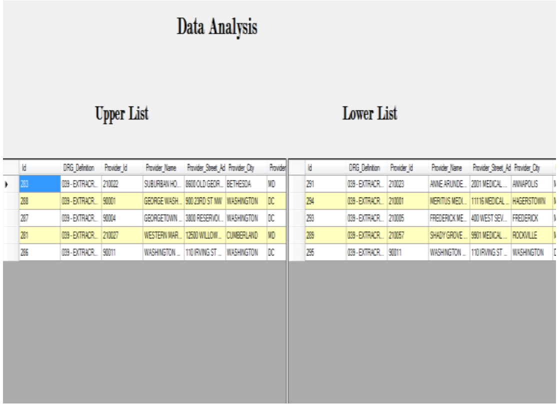
DRG Definition	Provider_Id	Provider_Name	Provider_Street_Addr	Provider City	Provider_State	Provider_Zip_Code	Hospital_Referral_Re	Total_Discharges	Total_Payments
039 - EXTRACRAN...	90001	GEORGE WASHIN...	900 23RD ST NW	WASHINGTON	DC	20037	DC - Washington	11	10582.81
039 - EXTRACRAN...	90004	GEORGETOWN U...	3800 RESERVOIR ...	WASHINGTON	DC	20007	DC - Washington	14	9479.57
039 - EXTRACRAN...	90011	WASHINGTON H...	110 IRVING ST NW	WASHINGTON	DC	20010	DC - Washington	22	8419.54
039 - EXTRACRAN...	210001	MERITUS MEDICA...	11116 MEDICAL C...	HAGERSTOWN	MD	21742	DC - Washington	28	6106.14
039 - EXTRACRAN...	210005	FREDERICK MEM...	400 WEST SEVEN...	FREDERICK	MD	21701	DC - Washington	16	7626.12
039 - EXTRACRAN...	210022	SUBURBAN HOSP...	8600 OLD GEORG...	BETHESDA	MD	20814	DC - Washington	42	11198.68
039 - EXTRACRAN...	210023	ANNE ARUNDEL ...	2001 MEDICAL PA...	ANNAPOLIS	MD	21401	DC - Washington	79	5813.86
039 - EXTRACRAN...	210027	WESTERN MARY...	12500 WILLOWBR...	CUMBERLAND	MD	21502	DC - Washington	41	8715.17
039 - EXTRACRAN...	210057	SHADY GROVE A...	9901 MEDICAL CT...	ROCKVILLE	MD	20850	DC - Washington	23	8146.34



D:\Amol\Mtech Projects\Priyanka M

Id	DRG Definition	Provider_Id	Provider_Name	Provider_Street_Ad	Provider City	Provider_State	Provider_Zip_Code	Hospital_Referral_F	Total_Discharges	Total_Payments
1	039 - EXTRACR...	90001	GEORGE WASH...	900 23RD ST NW	WASHINGTON	DC	20037	DC - Washington	11	10582.81
2	039 - EXTRACR...	90004	GEORGETOWN ...	3800 RESERVOI...	WASHINGTON	DC	20007	DC - Washington	14	9479.57
3	039 - EXTRACR...	90011	WASHINGTON ...	110 IRVING ST ...	WASHINGTON	DC	20010	DC - Washington	22	8419.54
4	039 - EXTRACR...	210001	MERITUS MEDI...	11116 MEDICAL ...	HAGERSTOWN	MD	21742	DC - Washington	28	6106.14
5	039 - EXTRACR...	210005	FREDERICK ME...	400 WEST SEV...	FREDERICK	MD	21701	DC - Washington	16	7626.12
6	039 - EXTRACR...	210022	SUBURBAN HO...	8600 OLD GEOR...	BETHESDA	MD	20814	DC - Washington	42	11198.68
7	039 - EXTRACR...	210023	ANNE ARUNDE...	2001 MEDICAL ...	ANNAPOLIS	MD	21401	DC - Washington	79	5813.86
8	039 - EXTRACR...	210027	WESTERN MAR...	12500 WILLOW...	CUMBERLAND	MD	21502	DC - Washington	41	8715.17
9	039 - EXTRACR...	210057	SHADY GROVE ...	9901 MEDICAL ...	ROCKVILLE	MD	20850	DC - Washington	23	8146.34

Average: 8454.25
Total: 76088.21
Mode: 6106.14
Std Deviation : 1714.29



Data Analysis

Upper List

Id	DRG_Definition	Provider_Id	Provider_Name	Provider_Street_Ad	Provider_City	Provider
283	039 - EXTRACR...	210022	SUBURBAN HO...	8600 OLD GEOR...	BETHESDA	MD
288	039 - EXTRACR...	90001	GEORGE WASH...	900 23RD ST NW	WASHINGTON	DC
287	039 - EXTRACR...	90004	GEORGETOWN ...	3800 RESERVOI...	WASHINGTON	DC
281	039 - EXTRACR...	210027	WESTERN MAR...	12500 WILLOW...	CUMBERLAND	MD
286	039 - EXTRACR...	90011	WASHINGTON ...	110 IRVING ST ...	WASHINGTON	DC

Lower List

Id	DRG_Definition	Provider_Id	Provider_Name	Provider_Street_Ad	Provider_City
291	039 - EXTRACR...	210023	ANNE ARUNDE...	2001 MEDICAL ...	ANNAPOLIS
294	039 - EXTRACR...	210001	MERITUS MEDI...	11116 MEDICAL ...	HAGERSTOWN
293	039 - EXTRACR...	210005	FREDERICK ME...	400 WEST SEV...	FREDERICK
289	039 - EXTRACR...	210057	SHADY GROVE ...	9901 MEDICAL ...	ROCKVILLE
295	039 - EXTRACR...	90011	WASHINGTON ...	110 IRVING ST ...	WASHINGTON

V. CONCLUSION

In this paper,we propose hybrid algorithm to outlier detection by introducing support vector data iteration. We demonstrate the data density by using Gaussian method,minimizing the noise ratio and searching time.Our proposed method provide enhancement of the performance i.,e efficiency with the real time datasets also comparative study of results on the basis of graphs.BR tree methods using for indexing providing microclassification of data and normalize the data.

REFERENCES

[1] Bo Liu, Yanshan Xiao, Philip S.Yu, Zhifeng Hao, and Longbing Cao, "An Efficient Approach for Outlier Detection with Imperfect Data Labels," IEEE Trans.Knowl.Data Eng., Vol. 26, No. 7, July 2014

[2] Chang-Dong Wang, Jian-Huang Lai, Dong Huang, and Wei-Shi Zheng,"SVStream: A Support Vector-Based Algorithm for Clustering Data Streams,"IEEE Trans.Knowl. Data Eng.,vol. 25, No. 6, June 2013

[3] Y. J. Lee, Y. R. Yeh, and Y. C. F. Wang, "Anomaly detection via online oversampling principal component analysis,"IEEE Trans.Knowl. Data Eng.,vol. 25, no.7, pp. 1460–1470, May 2012.

[4] Xiaoli Li, Chris P. Bowers, and Thorsten Schnier, "Classification of Energy Consumption in Buildings With Outlier Detection,"IEEE Trans.On Industrial Electronics, Vol. 57, No. 11, November 2010

[5] C.C.Aggarwal and P.S.Yu,"A survey of uncertain data algorithms and applications,"IEEE Trans. Knowl. Data Eng.,vol. 21, no. 5, pp. 609–623, May 2009.

[6] S.Y.Jiang and Q. B. An,"Clustering-based outlier detection method," in Proc.ICFSKD, Shandong, China, 2008, pp. 429–433.

[7] Jeen-Shing Wang and Jen-Chieh Chiang, "A Cluster Validity Measure With Outlier Detection for Support Vector Clustering," IEEE Trans On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 38, No. 1, February 2008

[8] Hui Xiong, Gaurav PandeyMichael Steinbach and Vipin Kumar "Enhancing Data Analysis with Noise Removal,"IEEE Trans.Knowl. Data Eng.,vol. 18, No. 3, March 2006

[9] V.J.Hodge and J. Austin, "A survey of outlier detection methodologies,"Artif Intell.Rev. vol . 22, no. 3, pp, 85–126, 2004.

[10] W. Jansen, "Authenticating Mobile Device Users Through Image Selection," in Data Security, 2004.

[11] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special issue on learning from imbalanced data sets," SIGKDD Explorations, vol. 6, no. 1, pp. 1–6, 2004.

[12] M. V. Joshi and V. Kumar, "CREDOS: Classification using ripple down structure (a case for rare classes*),"* in Proc. SIAM Conf. Data Min., 2004.

[13] J. Theller and D.M. Cai, "Resampling approach for anomaly detection in multispectral images," in Proc. SPIE, Orlando, FL, USA, 2003, pp. 230–240.

[14] R. Smith, A. Bivens, M. Embrechts, C. Palagiri, and B. Szymanski, "Clustering approaches for anomaly based intrusion detection," in *Proc. Intell. Eng. Syst. Artif. Neural Netw.*, 2002, pp. 579–584.

[15] Y. Batistakis, M. Halkidi, and M. Vazirgiannis, "Cluster validity methods: Part i," in *Proc. ACM SIGMOD Rec.*, vol. 31. New York, NY, USA, pp. 40–45, 2002.

[16] G. Fumera and F. Roli, "Cost-sensitive learning in support vector machines," in *Proc. Workshop Mach. Learn. Meth. Appl.*, 2002.

[17] Y. Lin, Y. Lee, and G. Wahba, "Support vector machine for classification in nonstandard situations," *Mach. Learn.*, vol. 46, no. 1–3, pp. 191–202, 2002.

[18] P. Chan and S. Stolfo, "Toward scalable learning with nonuniform class and cost distributions," in *Proc. ACM SIGKDD Int. Conf. KDD*, 1998, pp. 164–168.

[19] G. Nakhaeizadeh, U. Knoll, and B. Tausend, "Cost-sensitive pruning of decision trees," in *Proc. ECML*, Catania, Italy, 1994, pp. 383–386

[20] V. Barnett and T. Lewis, *Outliers in Statistical Data.* Chichester,U.K.: Wiley, 1994.

[21] D. M. Hawkins, Identification of Outliers. Chapman and Hall, Springer, 1980