Integrated Intelligent Research (IIR)
International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 02 December 2013, Page No.55-59
ISSN: 2278-2419

A Review Paper on Outlier Detection using Two-Phase SVM Classifiers with Cross Training Approach for Multi- Disease Diagnosis

Kalpit R. Chandpa[1], Jignasa N. Patel[2]

[1]Department of Computer Science and Information Technology,
[2]Shri S'ad Vidya Mandal Institute of Technology, Bharuch 392-001, Gujarat, India
Email: [1]kalpit.it2011@gmail.com, [2]jpatel.25@gmail.com

Abstract-Data mining techniques can be effectively used for major disease prediction and diagnosis. But we argue that diagnosis of some major diseases those are correlated with several small diseases and highly interrelated with each other such as TB and HIV and multi disease diagnosis is tough and challenging task. In Data mining, Outlier detection is useful technique for medical diagnosis. When availability of class labels is there, supervised outlier detection successfully used to detect rare and abnormal objects for disease diagnosis purpose. Classification techniques such as Naïve Bayesian classifier, SVM, Association rule mining, neural networks are used for outlier detection. SVM is one of the best classification techniques for outlier detection as there is no requirement of explicit statistical model for SVM and it avoids dimensionality problem and provides optimum solution for classification. Outlier detection using SVM can be possible with One class SVM and Two-class SVM. But still One class SVM suffers with non availability of accurate class labels and Two-class SVM is suffer when one of the class is under sampled and long training time. This paper provides survey of disease diagnosis using SVM and Outlier detection using SVM on different domains and we have identified some key challenges in this field and introduced a solution of two-phase Outlier detection approach combining both Two-class SVM and One class SVM for prediction and diagnosis of multiple diseases those are highly interrelated with each other using available class labels of both data sets and proposed cross training approach for both SVM classifiers for reduce training time of SVM and accuracy purpose.

Keywords— Multi-disease diagnosis, Outlier detection, SVM, One class SVM, Two-class SVM**.**

I. INTRODUCTION

In recent years, Data mining has been widely used in the area of medical analysis, disease diagnosis and healthcare, called medical data mining. Various data mining tasks such as classification, clustering, association rule mining, and outlier analysis [1] can be applied on different diseases data set for disease prediction and disease diagnosis. But still diagnosis of diseases those are highly interrelated with each other and correlated with several small diseases means multi disease diagnosis is tough and challenging task.Outlier detection aims to find patterns in data that do not conform to expected behavior [2]. It has a wide variety of applications such as intrusion detection in cyber security, fraud detection for credit cards, fault or damage detection, insurance or health care, military surveillance for enemy activities and medical diagnosis. Outlier detection in the medical and public health domains typically work with patient records. The data can have outliers due to several reasons such as abnormal patient condition or instrumentation errors or recording errors. Thus the outlier detection is a very critical problem in this domain and requires high degree of accuracy [2]. Instead of removing those outliers, if those outliers are stored and properly analyzed it may very helpful for medical practitioners for disease diagnosis and decision making in this domain. If suitable Outlier detection technique is applied on health dataset for disease diagnosis purpose, it may give hidden and useful results from the dataset. There are so many techniques for outlier detection, but if some class labels are available then supervised outlier detection means classification based approaches are successfully used for outlier detection. Supervised or Classification based outlier detection techniques operate under the general assumption that a classifier can be learnt from a given feature space that can distinguish between normal and outlier classes [3].

SVM [4], [5], [6] is the popular classification-based approaches in the data mining and machine learning communities. It is widely used to detect outliers due to these main advantages: a).It does not require an explicit statistical model. b). It Provides an optimum solution for classification by maximizing the margin of the decision boundary. c). It avoids the dimensionality problem [7]. Support Vector Machines has been applied as outlier detection in the one-class setting called One-class SVM [8] as well as in Two-class fashion called Two-class SVM[9].This paper mainly focused survey of techniques for Outlier detection using Support vector machines on different domains including both one class SVM and two-class SVM. We have also provided survey on SVM based disease diagnosis and presented some key challenges and introduced a solution for those challenges by combining approach using both Two-class SVM and One-class SVM on multiple disease datasets and proposed the cross training approach using available class labels of both datasets for diagnosis of diseases those are highly interrelated with each other and correlated with some other diseases for reducing training time of both classifiers and accuracy purpose.This paper organized as follows: In Section II, we have discussed data mining techniques used for disease diagnosis by other researchers and indentified key challenges. Section III provides how outlier detection technique is useful in the field of medical diagnosis. In Section IV, we have discussed how SVM is used as an outlier detection technique and presented limitations of several proposed SVM based outlier detection techniques by

other researchers and given key challenges. In Section V, we introduce proposed solution and in Section VI; we have given conclusion and future enhancement of our proposed methodology.

II. LITERATURE REVIEW

K. Rama Lakshmi et al [10] discussed that Major data mining techniques such as association rule mining, classification, clustering and outlier detection can be useful for predicting and diagnosis of major life threatening diseases but accuracy of technique and algorithm is depends on the type of dataset and so for accuracy purpose combination of algorithms and techniques is proposed for accurately predict and diagnosis diseases. Karanjit Singh et al. [11] surveyed papers related with outlier detection in different domains and provided survey on outlier detection techniques and outlier detection applications such as intrusion detection, Fraud detection, Inside Training detection, Medical and Public Health outlier detection, Industrial damage detection, Image processing, Outlier detection in sensor networks. The important thing they have highlighted is two major issues that arise in supervised outlier detection. First, the anomalous instances are few, as compared to the normal instances in the training data. Second, obtaining accurate and representative labels, especially for the outlier class is usually challenging. Dr. Shuchita Upadhyaya et al. [3] have provided a structured and comprehensive overview of the research on Classification Based Outlier Detection including various techniques as applicable to our area of research. They have identified key assumptions, which are used by the techniques to differentiate between normal and Outlier behaviour. When applying a given technique to a particular domain, these assumptions can be used as guidelines to assess the effectiveness of the technique in that domain.

The real time network outlier detection method in the wireless sensor networks to classify the sensor node data as local outlier or cluster outlier or network outlier using Standard Support Vector Machine is proposed by M. Syed Mohamed et al. [7].Suya Xu et al. [12] Proposed KNN-SVM techniques (Support Vector Machines based on K-Nearest Neighbor Algorithm) for Outlier Detection in Wireless Sensor Networks.It utilizes KNN techniques to reduce training samples' scale which can shorten training time and optimize time. Then it maps the samples into feature space by kernel function. Problem with this KNN-SVM technique is that if KNN algorithms delete some outliers. For future, other techniques can be used to reduce the training time of SVM.Xiaoqi Peng et al. [13] Presented an outlier detection method based on SVM.

A SVM model built by the clean sample set without outlier is used to predict the samples, when the error between the prediction-value and actual value exceeds the threshold; the sample is taken as an outlier, otherwise a normal one for detecting and removing the high dimensional nonlinear outlier sample from the practical copper-matte converting production data. We found that if accurate class labels must there to train SVM otherwise it suffers with accuracy problem. SVM based disease diagnosis for breast cancer using SPSS Clementine data mining tool analysed with different kernel functions and parameters of SVM for solving existing problems in traditional individual breast cancer diagnosis and to improve the work of medical practitioners in the diagnosis of breast cancer by Shang Gao et al. [14]. Esraa M. Hashem et al. [15] used SVM for classifying liver disease using two liver patients' datasets with different features combinations such as SGOT, SGPT and Alkaline Phosphates for liver disease diagnosis purpose and implemented using MATLAB. Edward Smart et al. [16] compared One-class SVM and Two-class SVM For detecting multiple faults in induction motor and proved one-class SVM better than two-class SVM for fault detection. The main thing they have highlighted is classification through two-class SVM performance can suffer when one of the classes is under sampled. Classification through one-class SVM performance can suffer if accurate class labels for one class are not there.

III. DATA MINING TECHNIQUES FOR DISEASE DIAGNOSIS

In today's era, Data Mining is becoming popular in healthcare field because there is a requirement of efficient analytical methodology for detecting unknown and hidden information in health data. Data Mining provides several benefits such as detection of the fraud in health insurance, availability of medical solution to the patients at lower cost, detection of causes of diseases and identification of medical treatment methods in healthcare industry. It also helps the healthcare researchers for decision making for efficient healthcare policies, constructing drug recommendation systems, developing health profiles of individuals etc. Following table represents some data mining techniques for disease diagnosis

TABLE I

Data Mining Techniques For Disease Diagnosis

Technique	Purpose
Naïve Bayes, k-NN, and Decision List algorithm [17]	Heart disease diagnosis
C4.5, Naïve Bayes and Back-Propagated Neural Network [18]	Breast cancer diagnosis
SVM[14]	Breast cancer diagnosis
SVM[15]	Liver disease diagnosis
Weighted KNN classifier[19]	Skin disease diagnosis
Fuzzy K-NN[20]	Thyroid disease
Multilayer Neural Network [21]	chest disease diagnosis
k-means clustering [22]	Alzheimer's disease diagnosis
Incremental SVM [23]	Cardiovascular disease

Key Challenges

- ✓ Multi-disease diagnosis.
- ✓ Diagnosis of diseases those are highly interrelated with each other.

International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 02 December 2013, Page No.55-59
ISSN: 2278-2419

✓ Diagnosis of diseases those are correlated with several other diseases.

✓ High degree of accuracy is must.

IV. OUTLIER DETECTION FOR MEDICAL DIAGNOSIS

Medical applications generate and collects large amount of complex data in terms of diseases, symptoms, patients details etc. The generated and collected data may contain unusual patterns which suggest abnormal disease conditions for further medical analysis and medical diagnosis. Outlier analysis can play very crucial role in case of disease diagnosis by detecting rare and abnormal diseases as outliers from collection of diseases dataset which can be helpful for effective patient care.The healthcare data typically consists of patient records which may have several different types of features such as patient age, blood group, weight. Outlier analysis can be useful for medical fraud detection, as well as identifying rare and abnormal patterns from the disease datasets In public health data, outlier detection techniques are widely used to detect anomalous patterns in patient medical records which could be symptoms of a new disease [2].

V. OUTLIER DETECTION USING SVM

Outlier detection can be done by using effective classification and machine learning technique Support Vector Machines (SVMs) in both the one-class and two-class scenarios. Such techniques use one class learning to train a classifier that contains the training data objects means a boundary. In case of learning complex regions Kernels, like radial basis function (RBF) kernel can be used. This technique decides if the test instance falls inside the learnt region for each test instance, if a test object falls inside the learnt region, it is declared as normal and outside instance is considered as an outlier.

A.**Support vector machine**

In machine learning, support vector machines (SVM) are supervised learning models with associated learning algorithms that analyse data and recognize patterns, used for classification and regression analysis [1]. When a set of training examples are given to the SVM classifier, each supposed as belonging to one of two classes, an SVM training algorithm builds a model that sets new examples into one category or the other. This makes it as a non-probabilistic binary linear classifier. These examples are represented as points in space by an SVM model and mapped in such a way that the examples of the different categories are separated by a clear gap. This gap is as wider as possible. Based on which side of the gap new examples are belongs, they mapped into that same space and predicted to fall into that category.

An SVM finds the best hyperplane that separates all data points or objects of one class from those of the other class classifies data for classifying the data. The meaning of best hyperplane is, one with the largest margin between the two classes is the best hyperplane for an SVM. Margin indicates the maximal width of the slab parallel to the hyperplane that has no interior data objects or points.The following figure shows the basic concept of SVM in which + indicates data points of type 1, and – indicates data points of type –1. The data points that are closest to the separating hyperplane are called support vectors. These points are on the boundary of the slab.

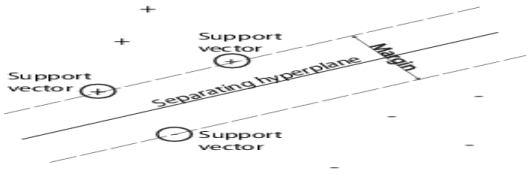

Figure 1: Basic Concept of SVM [6]

Outlier detection using SVM can be possible in following ways:

B. **Two-class SVM**

The two class SVM classifier seeks to maximise the margin between the healthy and faulty classes [16]. In the case of two-class SVM, classifier is trained according to the class labels of the two sets of data means normal and abnormal. When two-class SVM is used as outlier detection, output gives two classes, one is normal class and another is outlier class.

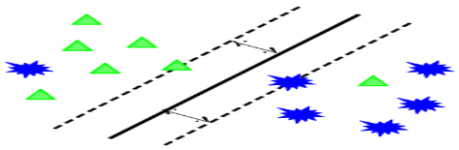

Figure 2: Basic concept of two-class SVM [24]

Performance of the two-class SVM can suffer when one of the classes is under- sampled [16]. Figure 2 represents basic concept of two-class SVM classifier.

C. **One-class SVM**

One class SVM classifiers are designed for novelty detection. In this case the aim is to separate a small number of abnormal data points from a large number of normal data points. They are designed to for situations where one class (the healthy class) is well sampled and the other class (the faulty class) is very poorly sampled [24].This makes it hard to use information from the poorly sampled class to determine a boundary between the two classes. One more problem with classification through one-class SVM is its performance can suffer if accurate class labels for one class are not there.

Figure 3 represents basic concept of two-class SVM classifier.

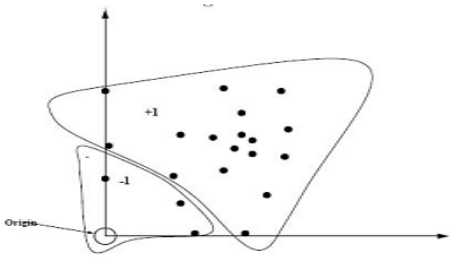

Figure 3: Basic concept of one-class SVM [24]

SVM based outlier detection applied on several domains for analysis purposes but still these approaches have some limitations and challenges. In following table, we presents SVM based outlier detection techniques with its purpose and

International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 02 December 2013, Page No.55-59
ISSN: 2278-2419

limitations and we have identified some key challenges in the field of Outlier detection using SVM.

TABLE II

A Survey Of Outlier Detection Using Svm

Algorithm	Purpose	Limitations
SVM [7]	To identify and classify outliers in the Wireless Sensor Network Real Time Data	Long training time for SVM, computational complexity
KNN-SVM[12]	To reduce training samples' scale this can shorten training time and optimize time and identifying outliers from Wireless Sensor Network Data efficiently.	KNN algorithms may delete some outliers.
SVM [13]	For detecting and removing the high dimensional nonlinear outlier sample from the practical copper-matte converting production data.	Availability of accurate class labels for clean samples is must, long training and testing time.
One-class SVM [25]	Identifying abnormal cases in the domain of melanoma prognosis.	Highly depends on accurate availability of class labels of one class.
Two-class SVM and One-class SVM [16]	For detecting multiple faults in induction motor using one-class and two-class SVM.	Two-class SVM performance can suffer when one of the classes is under sampled. One-class SVM performance can suffer if accurate class labels for one class are not there.
Two-class SVM and One-class SVM [24]	For detecting normal mammograms for breast cancer diagnosis.	Two-class SVM requires more space and time.

KEY CHALLENGES

- ✓ Techniques required for reducing SVM training time.
- ✓ Availability of accurate class labels is must.
- ✓ Applicability for multi-disease diagnosis.
- ✓ Minimizing computational complexity.

VI. PROPOSED SOLUTION

We proposed a solution of two-phase Outlier detection approach combining both Two-class SVM and One class SVM for prediction and diagnosis of multiple diseases those are highly interrelated with each other using available class labels of both data sets with cross training approach for both SVM classifiers for reduce training time of SVM and accuracy purpose.The main purpose of proposed system is to detect outliers in health dataset using SVM classifiers for disease diagnosis purpose. It includes following two phases:

❖ 1st Phase:

In the first phase outlier detection using two-class SVM classification is applied on both datasets D1 and D2. In this phase both classifiers C1 and C2 are trained according to class labels of both data sets means according to D1 and D2 respectively. After this phase 1, we get two classes for each data set normal class N1 and Outlier class O1.

❖ 2nd Phase:

In the second phase outlier detection using one class SVM classification is applied on both normal class and outlier class of both data sets. But this time classifier 2 means C2 is trained with D2 class label for D1 and D1 class label for D2. At the end of this process we will get again class with normal instances N2 and class with outlier instances O2.

VII. CONCLUSION AND FUTURE ENHANCEMENT

From the above discussion and survey, it is clear that data mining techniques are successfully used for particular single disease diagnosis, but some major diseases correlated with several small diseases and strongly interrelated with each other are exist such as TB and HIV, and diagnosis of these type of multi-diseases is critical issue and not much efforts are done using data mining techniques for solving this issue. This multi-disease diagnosis may very helpful for medical practitioners for analysis purpose and discovering hidden relationships among them.Our effort is to diagnosis of those major diseases using outlier detection based on SVM. For solving limitations of both two-class and one-class SVM, we have introduced a combine approach of two phase outlier detection and for solving issue of SVM training time and non availability of accurate class labels, we proposed a new cross training approach using available class labels of the both datasets for reducing time and accuracy purpose. For future, we will implement our proposed methodology on using some major diseases datasets and measured performance of our proposed methodology in terms of training time and diagnosis accuracy.

REFERENCES

[1] Jiawei. Han and Micheline Kamber, "Data Mining: Concepts and Techniques.pdf, 2nd Edition.
[2] Varun Chandola, Arindam Banerjee and Vipin Kumar, "Outlier Detection: A Survey.pdf"
[3] Dr. Shuchita Upadhyaya and Karanjit Singh, "2.3 Classification Based Outlier Detection Techniques", International Journal of Computer Trends and Technology- volume3, Issue 2- 2012.
[4] Wu, Shih-Hung, "Support Vector Machine Tutorial.pdf".
[5] Jason Weston, "Support Vector Machine and Statistical Learning Theory Tutorial.pdf".
[6] Chih-Wei Hsu, Chih-Chung Chang ,and Chih-Jen Lin, "A Practical Guide to Support Vector Classication.pdf".
[7] M. Syed Mohamed and T. Kavitha, "Outlier Detection Using Support Vector Machine in Wireless Sensor Network Real Time Data",

International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-2, May 2011.

[8] B.Scholkopf ,R.C.Williamson, A.J. Smola, J.Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," in Advances in Neural Information Processing Systems 12,S.A. Solla, T. K. Leen, and K.- R. Muller, Eds. Cambridge, MA: MIT Press, 2000, pp. 582–588.

[9] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers" in COLT'92: Proceedings of the fifth annual workshop on Computational learning theory. New York, NY, USA: ACM, 1992, pp.144–152.

[10] K. Rama Lakshmi and S.Prem Kumar, "Utilization of Data Mining Techniques for Prediction and Diagnosis of Major Life Threatening Diseases Survivability-Review", International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013.

[11] Karanjit Singh and Dr. Shuchita Upadhyaya, "Outlier Detection: Applications And Techniques", International Journal of Computer Science Issues, Volume 9, Issue 1, No-3, January-2013.

[12] Suya Xu, Caiping Hu, Lisong Wang, Guobin Zhang, "Support Vector Machines based on K-Nearest Neighbor Algorithm for Outlier Detection in WSNs", IEEE 2012.

[13] Xiaoqi Peng, Jun Chen, Hongyuan Shen," Outlier Detection Method Based on SVM and Its Application in Copper-matte Converting", IEEE 2010.

[14] Shang Gao and Hongmei Li, "Breast Cancer Diagnosis Based on Support Vector Machine", IEEE International Conference on Uncertainty Reasoning and Knowledge Engineering 2012.

[15] Esraa M. Hashem, Mai S. Mabrouk, "A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis" American Journal of Intelligent Systems 2014, 4(1): 9-14.

[16] Edward Smart, David Brown and Luke Axel-Berg, "Comparing One and Two Class Classification Methods for Multiple Fault Detection on an Induction Motor", IEEE Symposium on Industrial Electronics & Applications (ISIEA2013), September 22-25, 2013.

[17] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer science and Engineering, Vol. 3, No. 6, June 2011.

[18] Choi J.P., Han T.H. and Park R.W., "A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis", J Korean Soc Med Inform, 2009, pp. 49-57.

[19] C. Hattice and K. Metin, "A Diagnostic Software tool for Skin Diseases with Basic and Weighted K-NN", Innovations in Intelligent Systems and Applications (INISTA), 2012.

[20] D. Y. Liu, H. L. Chen, B. Yang, X. E. Lv, N. L. Li and J. Liu, "Design of an Enhanced Fuzzy k-nearest Neighbor Classifier Based Computer Aided Diagnostic System for Thyroid Disease", Journal of Medical System, Springer, 2012.

[21] O. Er, N. Yumusakc and F. Temurtas, "Chest diseases diagnosis using artificial neural networks", Expert Systems with Applications, vol. 37, (2010), pp. 7648-7655.

[22] J. Escudero, J. P. Zajicek and E. Ifeachor, "Early Detection and Characterization of Alzheimer's Disease in linical Scenarios Using Bioprofile Concepts and K-Means", 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA, (2011) August 30-September 3.

[23] Binod Kumar Mishra, Prashant Lakkadwala, Naveen Kumar Shrivastava, "Novel Approach to Predict CARDIOVASCULAR DISEASE using Incremental SVM ", IEEE International Conference on Communication Systems and Network Technologies 2013.

[24] Mona Y. Elshinawya, Abdel-Hameed A. Badwy, Wael W. AbdelMageed, Mohamed F. Chauikha, "Comparing One-class and Two-class SVM Classifiers for Normal Mamogram Detection.pdf "

[25] Stephan Dreiseit, Melanie Os, Christian Scheibbock Michael Binder, "AMIA Symposium 2010."