IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. VI (May-Jun. 2014), PP 78-83
www.iosrjournals.org

Detecting Outliers Using Classifier Techniques and Discovering Their Behaviours

Mrs. S. Nithya[1], Ms.M.R.Kavia[2]
Assistant professor in Department of Computer Science and Engineering[1,2]
Prist University, TamilNadu, India

Abstract: *An outlier is an observation that lies an abnormal distance from other values in a random sample from a given population. In given population consider the problem of discovering attributes or properties ,accounting for the previous stated abnormality of a group of anomalous individuals . we are using different types of classifiers like Support vector Mechanism, RBF Network, Bayes network etc to detect the outliers and find the behavior of outliers. For that we are using EXPREX Algorithm, for efficiently discovering exceptional properties and discover the attributes of which induces the outliers.*

Index Terms: *Knowledge Discovery, anomaly characterization, Machine Learning, Data Mining, WEKA, Classification*

I. Introduction

Information is provided that a (typically small) fraction of the individuals in that data population is anomalous, but no reason whatsoever is given as to why these individuals behave anomalously. Characterizing the behavior of such anomalous individuals and the work [1] precisely considers the problem of discovering attributes that account for the (a-priori stated) abnormality of one single individual within a given data population.

As an example, consider a rare disease and assume a population of healthy and unhealthy human individuals is given; here, it would be very useful to single out properties characterizing the unhealthy individuals. For that we have to find the exceptional property. For Using the Exceptional property first resort a form of Minimum Distance Estimation for evaluating the badness of fit values by outliers compared to the probability values by inliers.

Next Find the Exceptionality score. The score values may be numerical or either categorical. Scores are calculated by both of analytical and empirical point of view to detect the outliers. The categorical property can be tested by using Randomization test .It is Chi square Test. The Numerical Property we have to use Cramer – von – mises criterion. Also, we present an algorithm, called EXPREX, or **Exceptional Property EXtractor**, that automatically singles out the exceptional properties and their associated explanations. In order to make the significance of the kind of knowledge mined by the EXPREX algorithm clear, we briefly illustrate next a real life example application scenario.

II. Related Work

F. Angiulli et al. [1] in his work he explored ,the abnormality of combination of attribute values. Discovering the attributes and find the abnormality of individuals in a given dataset. They take Global and Local Proparties to find out the outliers. Global property is finding the abnormality of entire data population. Local Properties is taking two subsets of attributes are singled out in Global properties.

K. Pearson,[2] in his work he determined the goodness of fit in the given data to using Preliminary Proposition .

G. Passarino, et al [4] reported to take the dataset in old age in good health. They take Synthetic Survival Curve to separate the age between 18 to 106 years. Then take Multinomial Logistic Regression Model to evaluate genotypes. And Finally take the regression model to compare the cross sectional dataset.

L. U. Gerdes, et al [5] proposes the Genotype Distributions. Cross sectional studies of people used to differentiate the age of various groups. Then they apply the Apo lipoprotein e genotype (APOE). It is Used to estimate the mortality risk.So the result of APOE to find APOE gene and Frailty gene.

Xiong, H et al [13] in their paper Data cleaning methods focus on removing noise, result from an imperfect data collection process Traditional Outlier Detection Technique, Distance Based , Clustering Based ,Local Outlier Factor(LOF) ,Hyperclique-based data cleaner (Hcleaner). It leads to better clustering performance and higher

quality associations.

F. Angiulli et al. [8] proposed to find all subset of attributes examined outlying subset and outlying subspaces. It is very difficult to find subset because of exceptionsl growth. So its difficult to find outlying subspace too.So here using Outlying Subset Search Algorithm. For distance calculation they have to compute upper and lower bound.Here they found the goodness of fit value by using K – nearest Algorithm.Use Random Sampling Technique to reduce the computation of genetic algorithm. It is Unsupervised Frame that is used to identify abnormal instances in large complex semantic groups. Key attribute algorithm is used to find the outlying partition. To analyse the orgin used Theroy Rough set and Clustering Algorithm.

Pal et al [9] they explored Fuzzy-possibilistic c-means (FPCM) model – Unlabeled data ,Row sum constraint produces unrealistic typicality values for large data sets. Possibilistic-fuzzy c-means (PFCM) model.
It produces memberships and possibilities. Hybridization of possibilistic c-means (PCM) and fuzzy c-means (FCM). Eliminates the row sum constraints of FPCM. .

V. Hodge et al [10] described Outlier detection has been used for centuries to detect and, where appropriate, remove anomalous observations from data. Outliers arise due to mechanical faults, changes in system behaviour, fraudulent behaviour, human error, instrument error or simply through natural deviations in populations

D. Tax [12] explained Identify all objects which are neither apples nor pears. The problem in one-class classification is to make a description of a target set of objects and to detect which (new) objects resemble this training set. The difference with conventional classification is that in one-class classification only examples of one class are available. The objects from this class will be called the target objects. All other objects are per definition the outlier objects.

E. Knorr et al [13] proposed we present two simple algorithms, both having a complexity of O(k

N'), k being the dimensionality and N being the number of objects in the dataset. These algorithms readily support datasets with many more than two attributes. Second, we present an optimized cell-based algorithm that has a complexity that is linear wrt N, but exponential wrt k. Third, for datasets that are mainly disk-resident, we present another version of the cell-based algorithm that guarantees at most 3 passes over a dataset.

F. Angiulli et al [15] in this paper they explained a novel distance-based outlier detection algorithm, named DOLPHIN, working on disk-resident datasets and whose I/O cost corresponds to the cost of sequentially reading the input dataset file twice, is presented.

III. Methodology & Techniques

3.1Classification Techniques

3.1.1.Radial Basis Function(RBF)

The radial basis function (RBF) network is a special type of neural networks with several distinctive features . Since its first proposal, the RBF network has attracted a high degree of interest in research communities. A RBF network consists of three layers, namely the input layer, the hidden layer, and the output layer. The input layer broadcasts the coordinates of the input vector to each of the units in the hidden layer.

Each unit in the hidden layer then produces an activation based on the associated radial basis function. Finally, each unit in the output layer computes a linear combination of the activations of the hidden units. How a RBF network reacts to a given input stimulus is completely determined by the activation functions associated with the hidden units and the weights associated with the links between the hidden layer and the output layer.
Even when Bayesian methods are computationally intractable, they can provide a standard of optimal decision making against which other methods can be measured.

A Bayesian Network (BN) is a graphical model for probability relationships among a set of variables features. The Bayesian network structure S is a directed acyclic graph (DAG) and the nodes in S are in one-to-one correspondence with the features X. The arcs represent casual influences among the features while the *lack* of possible arcs in S encodes conditional independencies. Moreover, a feature (node) is conditionally independent from its non-descendants given its parents ($X1$ is conditionally independent from $X2$ given $X3$ if $P(X1|X2,X3)=P(X1|X3)$ for all possible values of $X1$, $X2$, $X3$).

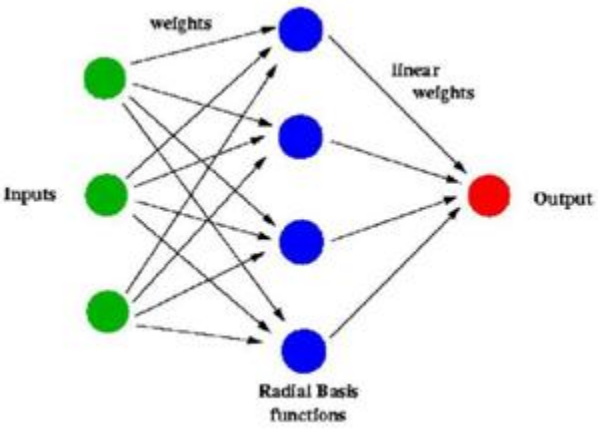

Fig :3.1.1 RBF Network

3.1.2. Bayes Network Classifier

Bayesian networks are a powerful probabilistic representation ,and their use for classification has received considerable attention. It's real statistical significance became much stronger. Each training example can incrementally increase/decrease the probability that a hypothesis is correct. Prior knowledge can be combined with observed data.

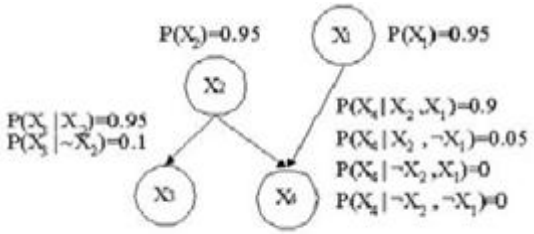

Fig 3.1.2 The Structure of Bayes Network

3.1.3. K-Nearest Neighbor Classifier

Nearest neighbor classifiers are based on learning by analogy. The training samples are described by n dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample. "Closeness" is defined in terms of Euclidean distance.

Nearest neighbor classifiers are instance-based or lazy learners in that they store all of the training samples and do not build a classifier until a new(unlabeled) sample needs to be classified. This contrasts with eager learning methods, such a decision tree induction and back propagation, which construct a generalization model before receiving new samples to classify.

3.1.4.Decision Tree and Pruning

Decision trees are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values.

There are two common approaches that decision tree induction algorithms can use to avoid over fitting training data: i) Stop the training algorithm before it reaches a point at which it perfectly fits the training data, ii) Prune the induced decision tree. If the two trees employ the same kind of tests and have the same prediction accuracy, the one with fewer leaves is usually preferred.

3.1.5. Single Conjunctive Rule Learner

Single conjunctive rule learner is one of the machine learning algorithms and is normally known as inductive learning. The goal of rule induction is generally to induce a set of rules from data that captures all generalizable knowledge within that data, and at the same time being as small as possible [6].

Classification in rule-induction classifiers is typically based on the firing of a rule on a test instance, triggered by matching feature values at the left-hand side of the rule [7]. Rules can be of various normal forms,

and are typically ordered; with ordered rules, the first rule that fires determines the classification outcome and halts the classification process.

3.1.6. Support vector Mechanism

Our algorithm maintains a candidate Support Vector set. It initializes the set with the closest pair of points from opposite classes like the Direct SVM algorithm. As soon as the algorithm finds a violating point in the dataset it greedily adds it to the candidate set. It may so happen that addition of the violating point as a Support Vector may be prevented by other candidate Support Vectors already present in the set. We simply prune away all such points from the candidate set.

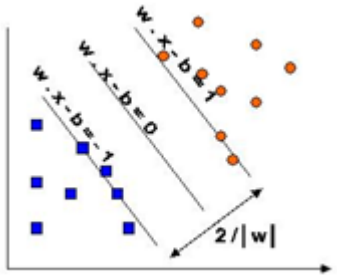

Fig 3.1.3 Maximum margin hyperplanes for a SVM trained with samples from two classes

3.1.7. Multi-layer Perceptron

Multi-layer perceptrons (MLP) are a popular form of feedforward artificial neural networks with many successful applications in data classification. The supervised learning (training) process of an MLP with input data x and target t, requires the use of an objective function (or error/cost/loss function) E(y; t) in order to assess the deviation of the predicted output values, y= MLP(X;W) from the observed data values t and use that assessment for the convergence towards an optimal set of weights w^*.

There are many MLP training algorithms using the əE/əW gradient information either directly or indirectly. In the present paper we concentrate on using the well-known back propagation (BP) algorithm without loss of generalization..

3.1.6.Other Classification

Like Association Rules ,Rough set Approach, Genetic Algorithm, Fuzzy set Approach are all classifier etc..
After Classification it gives the normal and abnormal datas. Abnormal Datas are known as Outliers . After find out the outliers we have to find the behavior of attributes of outliers.

IV. Outlier Detection

Using any one of the classifiers we can find the outliers of correctly classified and incorrectly classified outliers.

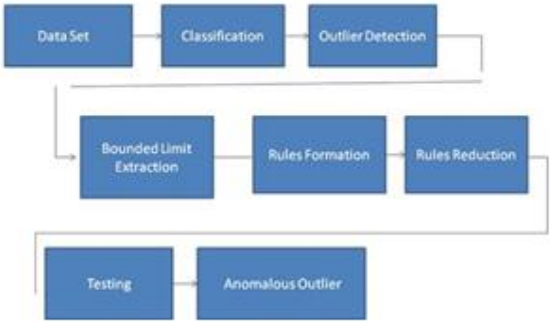

Fig 4. Module Diagram

V. Bounded Limit Extraction

After Classification we have see the inliers and outliers . First sort the both inliers and outliers by using Minimum Distance Algorithm and find the minimum value in both inliers and outliers Find the minimum value in both by using One to many . For that we use the algorithm of Generate Base Condition.

VI. Generate Base Condition

Find the Minimum value in both of inliers and outliers. Set the Outlier count. Calculate the Boundary Value by the range of outliers. Set the threshold value by Domain experts. Using EXPREX Algorithm to put all classifiers we have to find the attribute of outliers

Algorithm
1: EXPREX algorithm

Input: the outlier dataset *DBo* and the inlier dataset *DBi*

the outlier frequency threshold *_o* the inlier frequency threshold *_i*

Output: the exceptional explanation-property pairs

1: let *A* be the set of attributes of *DBo* and *DBi*

2: set *R* to *; // the set storing the exceptional explanation-property pairs*
3: **foreach** *a 2 A* **do**

// *base condition generation step* 4: *Ca = GenerateBaseConditions* (*DBo;DBi; a; _o; _i*)

5: **foreach** *p 2 A* **do**

// *base condition combination step*

6: let *B* be the whole set of conditions ∪ *a2*(*Anfpg*) *Ca*

7: *R = R*

[CombineBaseConditions(*DBo;DBi; p;B; _o; _i*) 8: **return** *R*

VII. Conclusion

This work aimed at providing a contribution towards the design of automatic methods for the discovery of properties characterizing a small group of outlier individuals as opposed to the whole population of"normal" individuals. In particular, we have introduced the concept of exceptional explanation-property pair and have discussed the significance of the associated knowledge.

References

[1]. F. Angiulli, F. Fassetti, and L. Palopoli,
[2]. "Detecting outlying properties of exceptional objects," ACM Trans. Database Syst., vol. 34, no. 1, 2009.
[3]. K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arised from random sampling," Philosophical Magazine Series 5, vol. 50, pp. 157–175, 1900.
[4]. T. W. Anderson, "On the distribution of the two-sample cram´ervon mises criterion," The Annals of Mathematical Statistics (Institute of Mathematical Statistics), vol. 33, no. 3, pp. 1148–1159, 1962.
[5]. G. Passarino, A. Montesanto, S. Dato, S. Giordano, F. Domma, V. Mari, E. Feraco, and G. D.
[6]. Benedictis, "Sex and age specificity of susceptibility genes modulating survival at old age," Human Heredity (Int. Journal of Human and Medical Genetics), vol. 62, no. 4, pp. 213–220, 2006.
[7]. L. U. Gerdes, B. Jeune, K. A. Ranberg, H. Nybo, and J. W. Vaupel, "Estimation of apolipoprotein e genotype-specific relative mortality risks from the distribution of genotypes in centenarians and middle-aged men: apolipoprotein e gene is a "frailty gene", not a "longevity gene"," Genetic Epidemiology, vol. 19, no. 3, pp. 202–210, 2000Cohen, W. (1995) Fast effective rule induction.
[8]. *In Press of Proceedings 12th International Conference on Machine Learning*, Morgan Kaufmann. Pp. 115–123.
[9]. Clark, P., Niblett, T. (1989). The CN2 rule induction algorithm. Machine Learning 3. pp. 261–284
[10]. Angiulli, F.; Basta, S.; Lodi, S.; Sartori, C.,"
[11]. Distributed Strategies for Mining Outliers in Large
[12]. Data sets" , IEEE Transactions on Knowledge and Data Engineering ,2003
[13]. Pal, N.R.; Pal, K.; Keller, J.M.; Bezdek, J.C, " A Possibilistic Fuzzy C - Means Clustering Algorithm " , IEEE Transactions on Fuzzy Systems ,2005
[14]. V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, 2004.
[15]. V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, 2004.
[16]. N. V. Chawla, N. Japkowicz, and A. Kotcz,
[17]. "Editorial: special issue on learning from imbalanced data sets," *SIGKDD Explorations*, vol. 6, no. 1, pp. 1–6, 2004.
[18]. Xiong, H.; Gaurav Pandey; Steinbach, M.; Vipin Kumar, "Enhancing data analysis with noise removal",IEEE Transactions on Knowledge and Data Engineering ,2006

[19]. .Nicolescu, M.; Medioni, G ," A Voting Based Computational framework for visual motion analysis and interpretation", IEEE Transactions on Pattern Analysis and machine Intelligence , 2005
[20]. D. Tax, "One-class classification," Ph.D. dissertation, Delft University of Technology, 2001.
[21]. A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/_mlearn/MLRepository.html