International Journal of Research
International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 08, August 2015
Available at http://internationaljournalofresearch.org

Implementation and classification of anomalous detection with varying parameters

Mukti[1] & Hari Singh[2]

Computer Science and Engineering, N.C. College of Engineering, Israna
Panipat, India
1 kundumukti@yahoo.in
[2] Harirawat@rediffmail.com

ABSTRACT –

Anomalous detection is an area that is extensively used in our daily lives an routine procedures. This involves the search for items or events which do not confirm to an expected pattern. The patterns thus detected are called anomalies and often translate to critical and actionable information in several application domains. In this paper we will extend the perspective of this approach in order to be able to deal with groups, or subpopulations, of anomalous individuals. 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.

INTRODUCTION:

An exceptional property is an attribute characterizing the abnormality of the given anomalous group (the outliers) with respect to the normal data population (the inliers). Moreover, each property can have associated a condition, also called explanation, whose aim is to single out a (significant) portion of the data for which the property is indeed characterizing anomalous subpopulations. In order to single out significant properties, we resort to minimum distance estimation methods that are statistical methods for fitting a mathematical model to data. To judge the quality of a property, we make use of exceptionality scores that are functions measuring the badness of fit of the values assumed by the outliers compared to the probability distribution associated with the values assumed by the inliers.

The exceptionality scores here defined are based on a randomization test using the Pearson chi-square criterion for categorical properties, and on the Cramer's-von-Misses criterion for numerical properties. These criteria evaluate the badness of fit of a probability distribution F compared to a sample set. In particular, we employ as reference distribution F the empirical distribution function associated with the population of inliers and, as the sample set, the population of outliers. We note that the proposed exceptionality scores are specifically designed for the task at hand, in which we compare a rare population with a large population of normal individuals. Also, this project focuses on an important algorithm, called EXPREX, [Exceptional Property Extractor] that automatically singles out the exceptional properties and their associated explanations.

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.. An exceptional property is an attribute characterizing the abnormality of the given anomalous group with respect to the normal data population. Moreover, each property can have associated a condition, also called explanation, whose aim is to single out a significant portion of the data for which the property is indeed characterizing anomalous subpopulations.

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.

ANOMALY DETECTION:

Anomaly detection is an important problem that has been researched within diverse research areas and application domains. Many anomaly detection techniques have been specifically developed for certain application domains, while others are more generic. This survey tries to provide a structured and comprehensive overview of the research on anomaly detection. We have grouped existing techniques into different categories based on the underlying approach adopted by each technique. For each category we have identified key assumptions, which are used by the techniques to differentiate between normal and anomalous 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. For each category, we provide a basic anomaly detection technique, and then show how the different existing techniques in that category are variants of the basic technique.

Subgroup discovery: The Subgroup Discovery Task (SDT) aims at finding an interesting subgroup of objects with common characteristics with respect to a given attribute value pair, called target variable .It was introduced in for categorical domains and, recently, extended to numerical domains. The SDT outputs a subgroup of individuals, identified as those individuals of the population satisfying a set of conditions, whose behaviour on the target attribute-value pair is different from the behaviour of the population taken as a whole.

Exceptional Explanation-Property pair:

Let Do (outliers) and Di (inliers) be two datasets on a set of attributes A, and let be an exceptionality score. A pair h E; pi, where E is a set of conditions on A and p is an attribute of A not occurring in E, is an exceptional explanation-property pair in Do w.r.t. Di (or, simply, exceptional pair) if and only if the attribute p is an exceptional property for DoE w.r.t. Di E according to , that is, if p(DoE ;DiE) > . In this case the attribute p is said to be an exceptional property and the value p(DoE ;DiE) is called exceptionality (score) of p (with explanation E).

The innovativeness of the approach has been illustrated by highlighting the substantial differences with related techniques. Moreover, we have defined the concept of exceptionality score, which measures the badness of fit of the values assumed by the outliers with respect to the probability distribution associated with the values assumed by the inliers. Suitable

exceptionality scores have been introduced for both numeric and categorical attributes In the year 2013, F. Angiulli in his work 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 Properties 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. In the year 2008, P. Filzmoser, R. Maronna 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, the concept of exceptional explanation-property pair has been introduced and has discussed the significance of the associated knowledge.. These scores have been shown, from both the analytical and the empirical point of view, to be effective for small samples, as outlier sets usually are. Thus, the method has been explicitly conceived to deal with rare subpopulations. In year 2013 Fabrizio Angiulli and Fabio Fassetti worked on Exploiting domain knowledge to detect outliers He present a novel definition of outlier whose aim is to embed an available domain knowledge in the process of discovering outliers. Specifically, given a background knowledge, encoded by means of a set of first-order rules, and a set of positive and negative examples, His approach aims at singling out the examples showing abnormal behaviour. The technique he proposed is unsupervised, since there are no examples of normal or abnormal behaviour, even if it has connections with supervised learning, since it is based on induction from examples. We provide a notion of compliance of a set of facts with respect to background knowledge and a set of examples, which is exploited to detect the examples that prevent to improve generalization of the induced hypothesis. By testing compliance with respect to both the direct and the dual concept, He is able to distinguish among three kinds of abnormalities, that are irregular, anomalous, and outlier observations. This allows to provide a finer characterization of the anomaly at hand and to single out subtle forms of anomalies. Moreover, he present both exact and approximate algorithms for mining abnormalities. In year 2010 Fabrizio Angiulli and Fabio Fassetti proposes a method for detecting distance-based outliers in data streams under the sliding window model. The novel notion of one-time outlier query is introduced in order to detect anomalies in the current window at arbitrary points in- time. Three algorithms are presented. The first algorithm exactly answers to outlier Queries, but has larger space requirements than the other two. The second algorithms derived from the exact one, reduces memory requirements and returns an approximate answer based on estimations with a statistical guarantee. The third algorithms a specialization of the approximate algorithm working with strictly fixed memory requirements. Accuracy properties and memory consumption of the algorithms have been theoretically assessed. Moreover experimental results have confirmed the effectiveness of the proposed approach and the good quality of the solutions. In the year 2004 V. Hodge worked on A survey of outlier detection methodologies. He 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, The algorithm incorporates buffer management and novel estimation and pruning techniques. He also present results of applying this

International Journal of Research
International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 08, August 2015
Available at http://internationaljournalofresearch.org

algorithm to sales data obtained from a large retailing company, which shows the effectiveness of the algorithm. The rules that they discover have one item in the consequent and a union of any number of items in the antecedent. he solve this problem by decomposing it into two sub problems: 1.finding all item sets, called large item sets that are present in at least s% of transactions. 2. Generating from each large item set, rules that use items from the large item set. In year 2001 S. D. Bay and M. J. Pazzani, they present the problem of mining contrast sets: Conjunctions of attributes and values that differ meaningfully in their distribution across groups. They provide a search algorithm for mining contrast sets with pruning rules that drastically reduce the computational complexity. Once the contrast sets are found, they post-process the results to present a subset that are surprising to the user given what we have already shown. they explicitly control the probability of Type I error (false positives) and guarantee a maximum error rate for the entire analysis by using Bonferroni corrections. instrument error or simply through natural deviations in populations In the year 1998 E. Knorr proposed Algorithms for mining distance-based outliers in large datasets, he 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 w.r.t. N, but exponential w.r.t. 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. In year 1993 Rakesh Aggrawal, Tomasz Imielinski,Arun Swami,works on a large database of customer transactions. Each transaction consists of items purchased by a customer in a visit. they present an efficient algorithm that generates all significant association rules between items in the database.

OBJECTIVE

1) To study different data sets.
2) To study different data sets, calculate the exceptionality score, which measures the badness of fit of the values assumed by the outliers with respect to the probability distribution associated with the values assumed by the inliers.
3) To perform the work, we will use an algorithm, called naïve bayes, which efficiently discovers exceptional explanation-property pairs.

METHODOLOGY AND FINDINGS

DATA SET - WEATHER

Day	Temperature	Outlook	Humidity	Windy	Play Golf?
07-05	hot	sunny	high	false	no
07-06	hot	sunny	high	true	no
07-07	hot	overcast	high	false	yes
07-09	cool	rain	normal	false	yes
07-10	cool	overcast	normal	true	yes
07-12	mild	sunny	high	false	no
07-14	cool	sunny	normal	false	yes
07-15	mild	rain	normal	false	yes
07-20	mild	sunny	normal	true	yes
07-21	mild	overcast	high	true	yes
07-22	hot	overcast	normal	false	yes
07-23	mild	rain	high	true	no
07-26	cool	rain	normal	true	no
07-30	mild	rain	high	false	yes

today	cool	sunny	normal	false	?
tomorrow	mild	sunny	normal	false	?

International Journal of Research
International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 08, August 2015
Available at http://internationaljournalofresearch.org

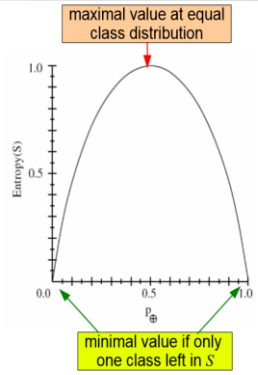
Entropy (for two classes)

- S is a set of examples
- $p_\oplus$ is the proportion of examples in class $\oplus$
- $p_\ominus = 1 - p_\oplus$ is the proportion of examples in class $\ominus$

Entropy:

$$E(S) = -p_\oplus \cdot \log_2 p_\oplus - p_\ominus \cdot \log_2 p_\ominus$$

- Interpretation:
  - amount of unorderedness in the class distribution of S



Example: Attribute Outlook

- Outlook = sunny: 3 examples yes, 2 examples no

$$E(\text{Outlook}=\text{sunny}) = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) = 0.971$$

- Outlook = overcast: 4 examples yes, 0 examples no

$$E(\text{Outlook}=\text{overcast}) = -1\log(1) - 0\log(0) = 0$$

Note: this is normally undefined. Here: = 0

- Outlook = rainy : 2 examples yes, 3 examples no

$$E(\text{Outlook}=\text{rainy}) = -\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) = 0.971$$

Entropy (for more classes)

- Entropy can be easily generalized for $n > 2$ classes
  - p_i is the proportion of examples in S that belong to the i-th class

$$E(S) = -p_1 \log p_1 - p_2 \log p_2 \ldots - p_n \log p_n = -\sum_{i=1}^{n} p_i \log p_i$$

Average Entropy / Information

Average entropy for attribute *Outlook*:

$$I(\text{Outlook}) = \frac{5}{14}\cdot 0.971 + \frac{4}{14}\cdot 0 + \frac{5}{14}\cdot 0.971 = 0.693$$

Information Gain

- When an attribute A splits the set S into subsets S_i
  - we compute the average entropy
  - and compare the sum to the entropy of the original set S

Information Gain for Attribute A

$$Gain(S, A) = E(S) - I(S, A) = E(S) - \sum_i \frac{|S_i|}{|S|} \cdot E(S_i)$$

- The attribute that maximizes the difference is selected

$Gain(S,\ Humidity\ )$	$Gain(S,\ Wind)$
= .940 - (7/14).985 - (7/14).592	= .940 - (8/14).811 - (6/14)1.0
= .151	= .048
$Gain(S, Outlook) = 0.246$	$Gain(S, Temperature) = 0.029$

Gain Ratio

Definition of Gain Ratio:

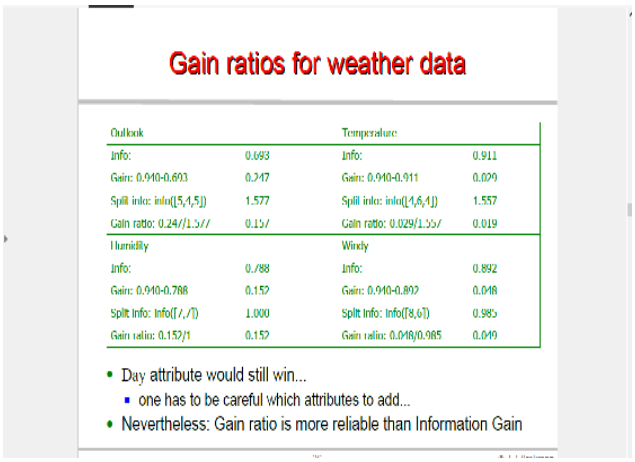
$$GR(S, A) = \frac{Gain(S, A)}{IntI(S, A)}$$

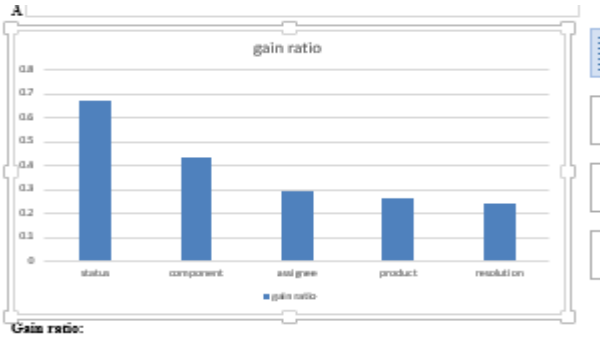
– .

Classification problems: require the prediction of a discrete target value

1. Can be solved using naïve Bayes.
2. Iteratively select the best attribute and split up the values according to this attribute.

International Journal of Research
International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 08, August 2015
Available at http://internationaljournalofresearch.org





Gain ratio:

In this graph we shows the gain ratio of every attribute. Here we find that status attribute has the max gain ratio and resolution attribute has minimum gain ratio. By gain ratio we resolve the data by removing data that is not of much use. So that we classify our data more accurately.

CONCLUSION

In this paper, we revisit some state-of-the-art naive Bayes text classifiers and empirically compare their classification performance on a large number of widely used classification benchmark datasets. Then, we propose a multinomial naive Bayes algorithm and its multiclass learning version called multiclass multinomial naive Bayes algorithm. The experimental results validate the effectiveness of our proposed algorithms. In future, we will customize NFR to improve the results for intrusion with reduced complexity and overheads and compare the efficiency rates with previous methods.

REFRENCES

[1]. Fabrizio Angiulli, Fabio Fassetti and Luigi Palopoli "Discovering Characterizations of the Behavior of Anomalous Sub populations" IEEE Transactions on knowledge and data engineering,vol. 25, no. 7, July 2012.

[2]. F. Angiulli, F. Fassetti, and L. Palopoli, "Detecting outlying properties of exceptional objects," ACM Trans. Database Syst., vol. 34, no. 1, 2009.

[3].H. Grosskreutz and S. Ruping, "On subgroup discovery in numerical domains," Data Mining and KnowledgeDiscovery, vol. 19, no. 2, pp. 210–226, 2009.

[4].F. Angiulli, R. Ben-Eliyahu-Zohary, and L. Palopoli, "Outlier detection using default reasoning," Artificial Intelligence (AIJ), vol. 72, no. 16–17,pp.1837–1872, November 2008.

[5]. F. Angiulli and C. Pizzuti, "Outlier mining in large high dimensional data sets,"IEEE Transactionon Knowledge and Data Engineering, vol. 17, no. 2, pp. 203–215, February 2005.

[6]. P. Filzmoser, R. Maronna, and M. Werner, "Outlier identification in high dimensions", Computational Statistics and Data Analysis, Volume 52, Issue 3, 1 January 2008,

[7]. V. Hodge and J. Austin, "A survey of outlier detection methodologies," Artif. Intell. Rev. vol. 22, no. 2, pp. 85–126, 2004.

[8]. E. Knorr and R. Ng, "Algorithms for mining distance-based outliers in large datasets," in Procs of VLDB-98, 1998, pp. 392–403.

[9].S. D. Bay and M. J. Pazzani, "Detecting change in categorical data: mining contrast sets," in KDD, 1999, pp. 302–306.

[10].S. D. Bay and M. J. Pazzani, "Detecting group differences: Mining contrast sets, "Data Mining and Knowledge Discovery, vol. 5, no., pp. 213–246, 2001.

[11].R. Agrawal, T. Imielnski, and A. Swami, "Mining association rules between sets of items in large databases," in SIGMOD. New York, NY, USA: ACM, 1993, pp. 207 216.

[12].S. Wrobel, "An algorithm for multi-relational discovery of subgroups," in PKDD, 1997, pp. 78–87.

[13]. K. Ramamohanarao, J. Bailey, and H. Fan, "Efficient mining of contrast patterns and their applications to classification," in ICISIP, 2005, pp. 39–47.

[14]. Shenchung dang "A fast Greedy Algorithm for outlier mining,"IEEE Transaction on Knowledge and Data Engineering, vol. 17, no. 2, pp.203–215, February 2006