Journal of Network and Innovative Computing (JNIC).
ISSN 2160-2174 Volume 1 (2013) pp. 23-32

Detecting Outliers in High Dimensional Categorical Data through Feature Selection

N N R Ranga Suri[1], M Narasimha Murty[2] and G Athithan[3]

[1]Centre for Artificial Intelligence and Robotics (CAIR),
C V Raman Nagar, Bangalore, India
rangasuri@gmail.com

[2]Department of CSA, Indian Institute of Science (IISc),
Bangalore, India
mnm@csa.iisc.ernet.in

[3]Centre for Artificial Intelligence and Robotics (CAIR),
C V Raman Nagar, Bangalore, India
athithan.g@gmail.com
Presently working at Scientific Analysis Group (SAG),
Delhi, India.

Abstract: **Extensive use of qualitative features for describing categorical data leads to high dimensional scenario in which outlier detection turns out to be a challenging task due to data sparseness. The curse of dimensionality has been well addressed in the case of numerical data by developing various feature selection methods, whereas the categorical data scenario is actively being explored. As the outlier detection problem is generally known to be unsupervised in nature due to lack of knowledge about various types of outliers, a novel unsupervised feature selection method is proposed in this paper for effective detection of outliers in categorical data. The proposed algorithm establishes the relevance and the redundancy of a feature through the entropy and the mutual information computation. By measuring the inherent redundancy of the features describing a data set, a threshold is applied on the allowed maximum redundancy of a candidate feature with already selected subset of features. This way of selecting features among the relevant ones results in a feature subset with less redundancy. The performance of the proposed algorithm in comparison with the information gain based feature selection shows its effectiveness for outlier detection. The efficacy of the proposed algorithm is demonstrated on various high-dimensional benchmark data sets employing two existing outlier detection methods.**
keywords: Data Mining, Outlier detection, Categorical data, Entropy, Mutual information

I. Introduction

The problem of outlier detection has been receiving a lot of attention in recent years due to its significance in dealing with various real life problems like fraud detection, anomaly detection, etc [1, 2, 3, 4]. A detailed discussion on various algorithmic issues involved in developing efficient data mining techniques for outlier detection has been brought out in [5]. While there exist many established methods [1, 2, 6, 7, 8, 9, 10] for detecting outliers in numerical data, there are only few methods like [11, 12, 13] that can process the data represented using categorical features/attributes. The problem of outlier detection in categorical data has been evolving, as evident from some recent publications [14, 15, 16] by various research groups.

The data pertaining to some of the application domains like market-basket analysis, biological data analysis tend to be very high dimensional in nature. Such data are typically described by a number of categorical features requiring methods that can scale well with the dimensionality. Data sparseness in high dimensional representation also makes the task of outlier detection more challenging. One can develop optimal algorithms for outlier detection by looking for outliers in sub-spaces of the data [17]. With increasing number of attributes (m) describing the data, one observes an exponential number ($2^m - 1$) of possible subspaces [8], making it practically an impossible task. In this context, identifying a useful sub-set of features aimed at outlier detection gains significance due to the presence of noisy and irrelevant features that degrade the performance of the detection process. As outlier detection is a well known class imbalanced problem with majority of the data objects belonging to the normal class and only a few objects being outliers, it turns out to be a more complex task. In such situations, it was suggested that resorting to feature selection is a necessary course of action [18]. In other such observation on highly imbalanced text classification problems, it was advocated that feature selection alone can combat the class imbalance problem [19].

A computational method based on random projection technique was proposed [20] recently, for outlier detection in high dimensional data. The objective of this method is to preserve the distances from various data objects to their k-nearest neighbors while projecting to a low-dimensional

space. Another approach for dealing with high-dimensional spaces applies eigenspace regularization on the training set in order to extract a relevant set of features for outlier detection [21]. More recently, a novel feature selection method applying a non-linear transformation in a feature-wise manner using kernel-based independence measures has been presented in [22]. Similarly, there are the other recent publications [9, 7, 8, 23] for detecting outliers based on the subspace clustering concepts. However, most of these methods are suitable primarily for numerical data.

A wrapper-based feature selection method was proposed in [24] for building classification models on binary class imbalanced data. Subsequently, a comparison of the methods developed for imbalanced data classification problems employing various features selection metrics was presented in [25]. Similarly, an approach employing feature bagging technique was proposed [26] by combining the results obtained from multiple instances of the outlier detection algorithm applied using different sub-sets of the features set. Applying the information theory principles, various mutual information based techniques have been proposed for feature selection [27, 28, 29] in the recent days. However, all these methods are applicable mainly for supervised learning problems. The Laplacian score-based method proposed in [30] and the feature similarity based method presented in [31] perform feature subset selection in unsupervised manner. Similarly, a novel feature selection method was proposed recently for dealing with the problems having multi-cluster data [32]. Though these methods can be employed for unsupervised learning tasks, their computation deals with only numerical data and their performance on class imbalanced data needs to be explored.

As brought out in [3], outlier detection is generally considered as an unsupervised learning problem due to lack of prior knowledge about the nature of various outlier objects. Moreover, unlabeled data is available in abundance, while obtaining labeled data is expensive in most of the applications. This motivates the need to develop an efficient unsupervised method for selecting features of relevance for outlier detection in categorical data. It is in this context, we propose a novel unsupervised feature selection algorithm employing mutual information measure for characterizing the redundancy among various features. This is basically a filtering technique that establishes the utility of a categorical feature in accomplishing the outlier detection task through feature-wise entropy computation. Features thus selected are expected to highlight the deviations characterizing the outliers with minimum redundancy among them. It is important to note that wrapper methods cannot be considered in the context of outlier detection due to the unsupervised learning requirement envisaged above. The experimental results furnished in [33] indicate the technical merit of the proposed method. Further theoretical discussion and additional experimental results on the proposed method are included in this paper for strengthening the claims regarding this method.

The rest of this paper is organized in four sections. Section II provides a quick view of the basics of the information theory with emphasis on the mutual information measure for feature selection. The subsequent section covers some mathematical preliminaries and the proposed feature selection algorithm along with some of its important properties, followed by Section IV giving various details of the experimental evaluation of the proposed algorithm such as the feature selection process and its impact on the outlier detection performance on various benchmark data sets. This section also includes a comparative study of the performance of the proposed method with that of an established supervised feature selection method. Finally, Section V concludes this paper with some discussion and directions for further work.

II. Related Work

According to the information theory, entropy measures the uncertainty associated with a random variable. The entropy $H(x)$ of a discrete random variable x is given by

$$H(x) = -\sum_{j} p(x_j)\log(p(x_j)) \tag{1}$$

where $p(x)$ is the marginal probability of x. Similarly, Mutual Information (MI) corresponds to the interdependence between two random variables x and y, defined as

$$I(x,y) = H(x) + H(y) - H(x,y)$$

For any two discrete random variables x and y, the MI value can be computed as

$$I(x,y) = \sum_{i}\sum_{j} p(x_i,y_j) log \frac{p(x_i,y_j)}{p(x_i)p(y_j)} \tag{2}$$

where the joint probability distribution $p(x,y)$ and the marginal probability distributions $p(x)$, $p(y)$ are determined by counting method.

A. Feature Selection using Mutual Information

Feature selection is the process of identifying a subset of the given features according to some criteria for removing irrelevant, redundant or noisy features. This has been an active research problem in data mining as it speeds up the learning process through dimensionality reduction and improves the learning accuracy by removing the noisy features.

Applying MI as a measure of relevance and redundancy among features for feature selection was initially proposed in [34]. As computing the joint MI between a multi-dimensional feature vector ($\{f_i, f_j\}$) and the class variable (C) is impractical, only $I(C;f_i)$ and $I(f_i;f_j)$ are computed. According to this method, MI measures the information content of a given feature with regard to the learning task at hand. Subsequently, various MI-based methods for feature selection have been proposed in the recent past for accomplishing various learning tasks. The minimal-redundancy-maximal-relevance (mRMR) criterion-based method proposed in [27] performs feature selection in an incremental manner by selecting one feature at a time. Having selected a set S_{k-1} of $(k-1)$ features, the one that maximizes the following expression is selected next as the k^{th} one.

$$G = [I(f_i;C) - \frac{1}{k-1}\sum_{f_s \in S_{k-1}} I(f_i;f_s)] \tag{3}$$

In the subsequent efforts, an improved method named Normalized Mutual Information Feature Selection (NMIFS) was

proposed in [28]. Based on the observation that the mutual information between two random variables is bounded above by the minimum of their entropy values, the NMIFS method makes use of the normalized MI value computed as

$$NMI(f_i, f_j) = \frac{I(f_i, f_j)}{min\{H(f_i), H(f_j)\}} \quad (4)$$

The normalization compensates for the MI bias toward multivalued features, and restricts its values to the range [0,1]. Accordingly, the selection criterion utilizes the average normalized MI between the candidate feature and the set of already selected features.

Referring to the details in [35, 28], MI has two distinguishing properties in comparison to other dependency measures: (i) the capacity of measuring any kind of relationship between two random variables and (ii) its invariance under space transformations, which are invertible and differentiable such as translations, rotations, and any transformation that preserves the order of the original elements of the variables. Thus, MI turns out to be an attractive option for feature selection, as each feature is considered to be a random variable. As the mRMR method takes the difference of the relevance term and the redundancy term (Equation 3), it may so happen that a redundant feature having relatively large relevance gets selected as one of the top features. Though imposing a greater penalty on the redundancy term would lessen this problem, it cannot be avoided completely. An alternative approach is to examine the relevance and the redundancy of a feature independently and establish its utility for the learning task at hand. This is the philosophy followed in the novel feature selection method proposed in this paper for accomplishing outlier detection.

III. An Unsupervised Feature Selection Scheme

Based on the theoretical foundations outlined above, we propose a novel unsupervised feature selection method leading to efficient outlier detection in high dimensional categorical data. Let the input to the proposed algorithm be an m-dimensional data set D consisting of n data objects with the descriptive features $F = \{f_1, f_2, ..., f_m\}$. The objective is to determine a suitable feature subset $F_s \subset F$ with as low redundancy as possible among the selected features, so as to detect the outliers present in the input data in an efficient manner. Accordingly, the outlier detection method employed here consists of two major tasks as shown in Figure 1.

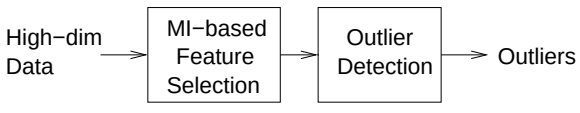


Figure. 1: Outlier detection through feature selection

The first task is to apply the proposed feature selection algorithm on the input data to determine a relevant subset of features F_s. As a result of feature selection, a low-dimensional version of the input data is produced, by keeping the values of the objects corresponding to the selected features. The next task is to apply a suitable outlier detection algorithm on this low-dimensional data for establishing the utility of the selected subset of features.

Two existing methods namely, the AVF method [11] and the Greedy method [12] have been considered here for outlier detection. The AVF method computes the frequency score of an object x_i as

$$AVFScore(x_i) = \frac{1}{m}\sum_{j=1}^{m} Freq(x_{ij})$$

where $Freq(x_{ij})$ is the number of times the j^{th} attribute/feature value of the object x_i appears in the data set. A lower score means that the object is more likely an outlier. Similarly, the Greedy algorithm proposed in [12] works based on the computation of the entropy of a data set given by

$$E(X) = E(X_1) + E(X_2) + E(X_3) + \cdots + E(X_m)$$

where $E(X_i)$ indicates the entropy value computed over the data set D corresponding to the feature f_i. According to this algorithm, a data object with maximum contribution to the entropy value gets labeled as the first outlier. It goes on identifying the outliers in successive iterations. In practice, any other outlier detection method working with categorical values can be employed here.

A. Mathematical Preliminaries

For the purpose of describing the proposed feature selection algorithm, we first introduce some mathematical notation utilized in this work.

Definition 1 *The redundancy of a feature f_i w.r.t. another feature f_j, denoted as $R(f_i, f_j)$, is the normalized mutual information (NMI) between these two features given as*

$$R(f_i, f_j) = NMI(f_i, f_j) \quad (5)$$

Definition 2 *The average redundancy of a feature f_i w.r.t. a set of features F, denoted as $AR(f_i, F)$, is defined to be the average of the pair-wise redundancy values of f_i w.r.t. each feature $f_j \in F$ as*

$$AR(f_i, F) = \frac{1}{|F|}\sum_{\forall f_j \in F} R(f_i, f_j) \quad (6)$$

Definition 3 *The average redundancy of a set of features F, denoted as $ALL\_AR(F)$, is defined to be the average of the average redundancy values of every feature $f_i \in F$ w.r.t. to the subset $F - \{f_i\}$ as*

$$ALL\_AR(F) = \frac{1}{|F|}\sum_{\forall f_i \in F} AR(f_i, F - \{f_i\}) \quad (7)$$

B. Proposed Algorithm for Feature Selection

Given the input data with m categorical features, the first step is to determine the feature-wise entropy values. A feature selection algorithm which produces a feature subset with less redundancy among the selected features is preferable. Accordingly, the proposed algorithm is designed to identify a subset of less redundant features by applying a threshold

value ($thresh$) for the allowed maximum redundancy among the selected features. Hence, determining a suitable value for this parameter in turn determines the qualifying features. The average redundancy value over all the features can be considered as a reasonable indicator of the intrinsic characteristic of the data and thus can be used as a reference value for the threshold selection. Accordingly, the next step in this process is to compute the average redundancy among the given set F of m features.

As mentioned in the previous section, the novel algorithm being proposed here is designed to deal with the relevance check and the redundancy evaluation of a feature independently. Accordingly, establishing the relevance of a feature for outlier detection is considered as the priority task over redundancy evaluation. Thus, we first arrange the set of features in non-decreasing order of their entropy values and designate the sorted list as F'. The rationale behind this process is that a feature contributing more to the detection of outliers (relevance) tends to have less entropy value (more skewed distribution of the values that it takes) as opposed to a feature primarily characterizing the normal objects. Though a similar situation may arise when the data set is having an attribute value with lower count without having any outliers, it doesn't occur naturally.

Having addressed the relevance aspect, the next task is to perform redundancy evaluation of a feature appearing in the relevance based ranked list. Following the heuristic defined for characterizing the relevance of a feature, features in the sorted list are considered one-by-one starting from the most relevant one, and the average redundancy value of each such feature $f'_i \in F'$ with respect to the selected subset F_s at that stage is determined using Equation 6. If this redundancy value happens to be less than the predefined threshold value, then that particular feature is included in the selected subset. Thus, the feature selection process progresses in an incremental manner till all the features in the sorted list F' are exhausted. The computational steps involved in the proposed feature selection procedure are listed in Algorithm 1.

Algorithm 1 A novel algorithm for unsupervised feature selection using Mutual Information (MI).

Input: A data set D with n objects and m descriptive attributes $F = \{f_1, f_2, ..., f_m\}$.
Output: Selected subset of features $F_s \subset F$.

1: Compute entropy $H(f_i)$ of each feature $f_i \in F$ as given in Equation 1.
2: Compute the average redundancy $ALL\_AR(F)$ as defined using Equation 7.
3: Set $thresh \leftarrow ALL\_AR(F)$.
4: Obtain the sorted sequence $F' = \{f'_1, f'_2, ..., f'_m\}$ in ascending order of the entropy values.
5: Initialize $F_s \leftarrow \{f'_1\}$, $i \leftarrow 2$
6: **while** $i \leq m$ **do**
7: **if** $H(f'_i) \neq 0$ **then**
8: Compute $AR(f'_i, F_s)$ using Equation 6.
9: **if** $AR(f'_i, F_s) \leq thresh$ **then**
10: $F_s \leftarrow F_s \cup \{f'_i\}$
11: **end if**
12: **end if**
13: **end while**

C. Analysis of the Proposed Algorithm

Depending on the value set for the parameter $thresh$, the allowed maximum redundancy among the selected features, a subset F_s of k features (where $k < m$) can be selected as per the computational steps listed in Algorithm 1. Based on the application requirement, a suitable value for this parameter can be determined resulting in a required number of selected features.

To simplify the analysis, we assume that every feature/attribute has a constant number of distinct values. Then, the marginal probability values can be computed with a single scan of the input data in $O(nm)$ computations employing hash tables. Similarly, the feature pair-wise joint probability values can also be computed. Then, the entropy and mutual information values can be determined using these probability values. All the remaining steps in the proposed algorithm can be carried out without further reading of the input data. Thus, the time complexity of the proposed algorithm turns out to be linear in terms of n, making it suitable for feature selection in large data sets.

Corollary 1 *The average redundancy of a feature subset F_s obtained using the steps in Algorithm 1, will be at most the value set for the threshold parameter.*

The Corollary 1 holds true as each selected feature in F_s has its average redundancy within the threshold value and the average of such values again turns out to be within the threshold value.

The following are some note worthy properties of the proposed feature selection algorithm:

- *Decoupling relevance from redundancy*: Unlike some existing algorithms, the relevance check and the redundancy evaluation of a candidate feature are carried out independently.
- *Relevance takes priority over redundancy*: Determining the relevance of a feature is attempted before evaluating its redundancy with emphasis on the intended learning task.
- *Deterministic relevance/redundancy computation*: Unlike the situation in the case of continuous valued features, mutual information computation on discrete features doesn't require any density estimation. To that extent, the computations involved here are deterministic and repeatable in nature.

IV. Experimental Evaluation

As brought in the proposed scheme shown in Figure 1, an experimental evaluation is carried out to demonstrate the efficacy of the proposed algorithm for feature selection and its impact on outlier detection performed on the resultant low dimensional representation of the data.

A. Details of the Benchmark Data Sets

The proposed method has been evaluated on some real life data sets taken from the UCI ML Repository [36]. As the

objective is to establish the usefulness of the proposed unsupervised feature selection algorithm for outlier detection, data sets with reasonably high dimensionality are better candidates for consideration. Accordingly, six categorical data sets haven been chosen for this experimentation as described in Table 1.

Data Set Name	# Features	# Objects	# Classes
Mushroom	22	8124	2
Chess (KR/KP)	36	3196	2
Splice-Junction	61	3190	3
Lymphography	18	148	4
Promoters Genes	58	106	2
SPECT Heart	22	80	2

Table 1: Details of the benchmark data sets

As per the standard practice in this field, objects with missing feature values have been eliminated. Corresponding to each one of the data sets, the minority (least sized) class objects were designated as outliers and all the remaining objects as normal ones, as shown in Table 2. Though these designated outliers are not outliers in real sense, they are considered so for validating the proposed method. To induce class imbalance in the data, only a subset (every 5^{th} object) of the designated outlier class objects have been considered across all data sets. Lymphography data is the only exception to this object subset selection as this data set has only six objects corresponding to the designated outlier class(es). In case of Promoters genes data and Splice-Junction data, the instance name field (second column) has been removed in pre-processing, resulting in 57 and 60 features respectively. As the proposed algorithm works in unsupervised learning mode, it doesn't require labeled data. However, class labels are used to measure its performance in detecting outliers.

B. Feature Selection Results

According to the proposed scheme, feature selection has been performed as the first task by applying the proposed unsupervised algorithm on various benchmark data sets considered in this experimentation.
Figure 2 shows the ranked sequence of features in ascending order of their relevance ranks, with the most relevant feature being the first ranked feature. Each feature in this ranked sequence is examined for its possible inclusion in the selected set of features as per the redundancy criterion defined in the proposed algorithm. All such selected features in the ranked sequence are marked with a '*' symbol on the corresponding impulse representation. It is important to note that corresponding to two data sets shown in Figure 2(a) and (b), the most relevant feature was not selected as its relevance (entropy) value was found to be '0', though it appeared first in the ascending rank sequence of features in both the cases. Looking at the feature/attribute values of those two data sets, it was found that this particular feature was having a single value through out the data set resulting in '0' entropy value. Such a feature is anyway not a good candidate for the discrimination of outliers from the normal objects, hence dropped by the feature selection algorithm.
As mentioned in the previous section, a threshold value ($thresh$) on the allowed maximum redundancy among the selected set of features determines the suitability of a feature for selection. Accordingly, the details on the average redundancy values and the chosen threshold values corresponding to the benchmark data sets are furnished in Table 3 for a better understanding.

Data Set Name	Normal Class	# Normal Objects	Outlier Class	# Outlier Objects
Mushroom	e	4208	p	783
Chess (KR/KP)	won	1669	nowin	305
Splice-Junction	EI,IE	1535	N	331
Lymphography	2,3	142	1,4	6
Promoters Genes	+	53	-	10
SPECT Heart	0	40	1	8

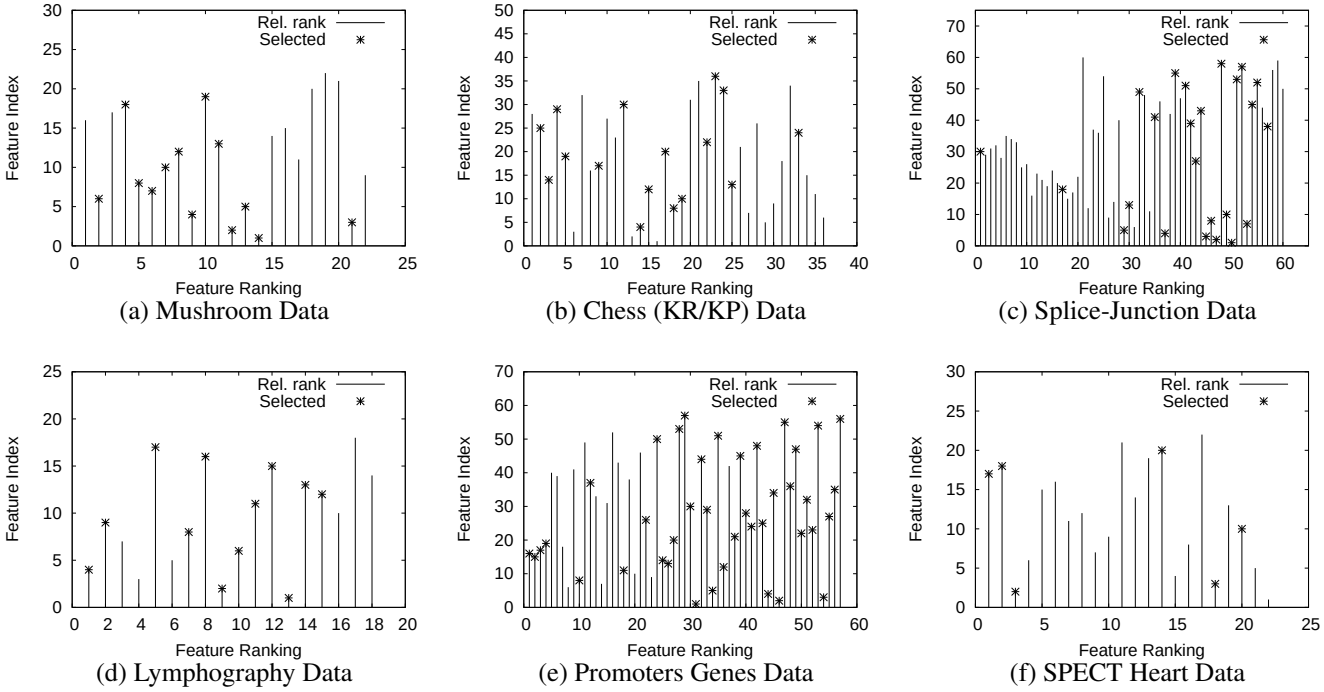
Table 2: Details of normal and outlier objects

Data Set Name	# Features (original)	Average Red.	Red. Thresh.	# Features (selected)
Mushroom	22	0.236	0.236	13
Chess (KR/KP)	36	0.072	0.03	16
Splice-Junction	60	0.01	0.008	24
Lymphography	18	0.087	0.1	12
Promoters Genes	58	0.072	0.09	41
SPECT Heart	22	0.103	0.1	6

Table 3: Redundancy threshold values set

C. Outlier Detection Results

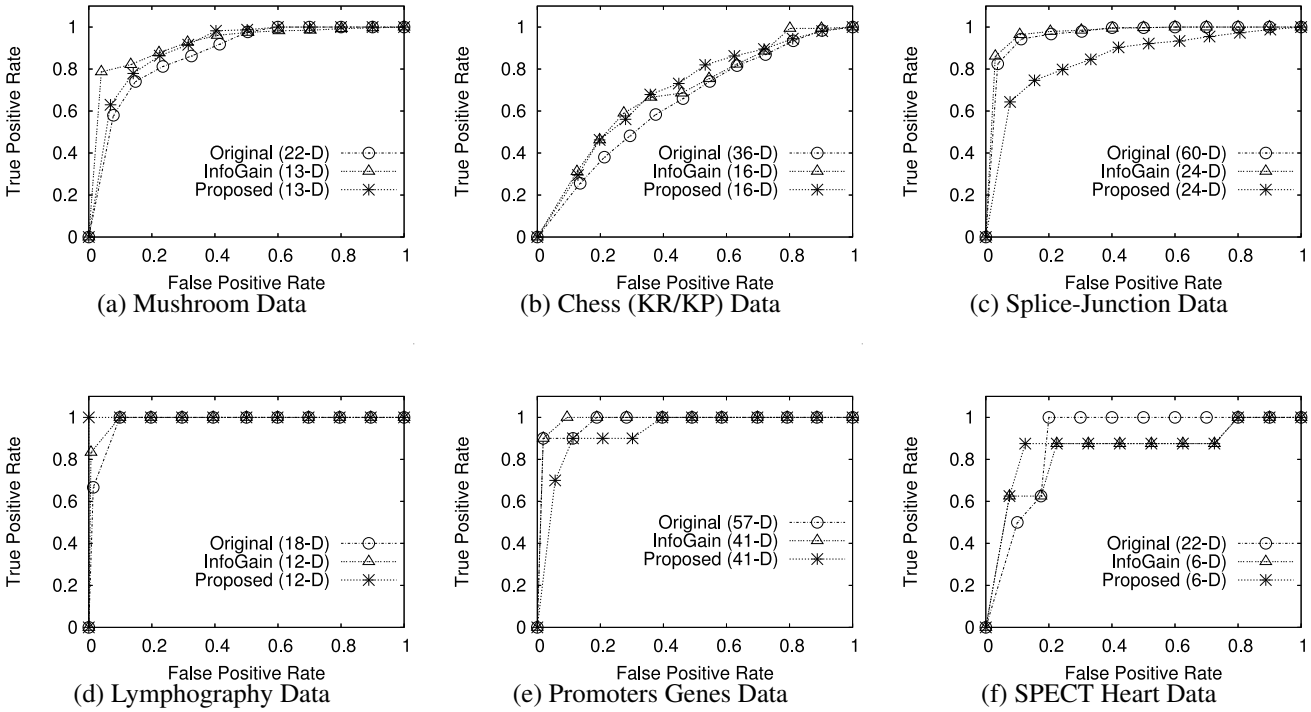
After performing feature selection on each one of the above considered data sets, the resultant low dimensional data sets were subjected to outlier detection using the AVF method and the Greedy method discussed in the previous section. For the purpose of comparing with the performance of the proposed algorithm, the information gain (IG) based feature selection method has been considered here, as it is shown to be one of the best performing methods for feature selection in class imbalanced problems [25] like outlier detection. Since the IG-based selection is performed in supervised setting, it is considered as the baseline method in this evaluation. Accordingly, outlier detection was carried out on the low dimensional data sets obtained using the IG-based selection as well, along with the proposed unsupervised method.
As pointed out in [3], there exists a trade off between the detection rate and detection accuracy in the context of outlier detection due to class imbalance. Following the general practice in such application contexts, we report the performance of the proposed method using the ROC curves [37]. Accordingly, the true positive rate (TPR) values have been determined with increasing false positive rate (FPR) values on every data set. The ROC curves thus generated applying the AVF method on the low dimensional data (obtained using the proposed and the baseline methods using the same number of selected features) as well as on the original data have been furnished in Figure 3 for a quick comparison. Similarly, Figure 4 depicts the outlier detection results obtained employing the Greedy method. In both the cases, the effectiveness of the proposed feature selection algorithm is evident from the superior or comparable performance obtained here with the reduced set of features vis-a-via the full set.



(a) Mushroom Data (b) Chess (KR/KP) Data (c) Splice-Junction Data

(d) Lymphography Data (e) Promoters Genes Data (f) SPECT Heart Data

Figure. 2: Ranked sequence of features according to their relevance with their selection status



(a) Mushroom Data (b) Chess (KR/KP) Data (c) Splice-Junction Data

(d) Lymphography Data (e) Promoters Genes Data (f) SPECT Heart Data

Figure. 3: Effect of feature selection on outlier detection performance (using the AVF method)

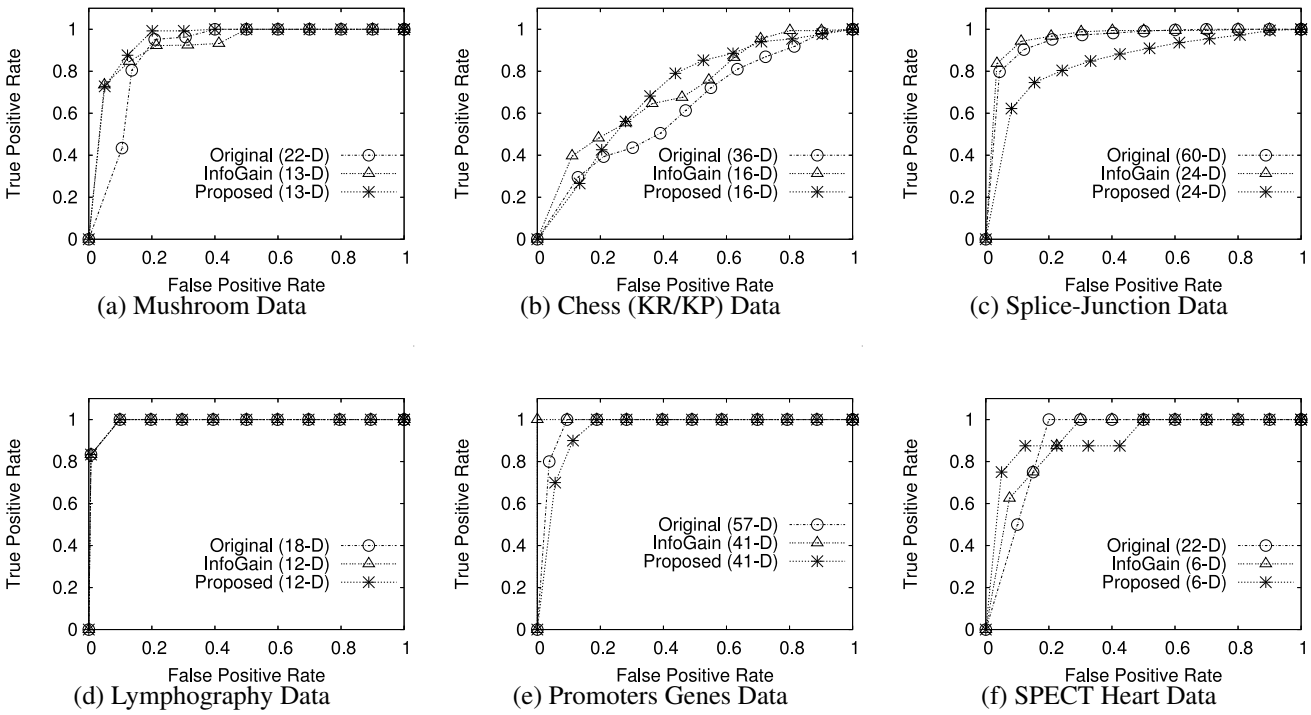


Figure. 4: Effect of feature selection on outlier detection performance (using the Greedy method)

D. Effect of Number of Features Selected

An issue of interest in this experimentation is to understand the effect of the number of features selected on the performance of the outlier detection process. To study this effect, we considered a micro array gene expression data set, namely the Lung cancer data, which is of high dimensionality (> 100 features). This data was discretized (as 3-states) and provided by [27] after pre-processing. It consists of 73 data objects belonging to 7 different classes described using 325 attributes. Similar to the process described above, data objects corresponding to two high sized classes (classes 7 and 4) were designated as normal objects and the objects belonging to one small sized class (class 3) as outliers. The effectiveness of the proposed feature selection algorithm on this data set is shown in Figure 5. Further experimentation has been carried out by varying the redundancy threshold parameter of the proposed algorithm. The detection accuracies corresponding to various cardinalities of the selected feature sub-sets are shown in Figure 6. Referring to this figure, it is quite clear that more number of outliers were detected corresponding to feature subset cardinalities between 40 and 250 than with the original feature set of 325 and the known highest was obtained with 41 features suggesting that a suitable low-dimensional representation can yield improved outlier detection performance.

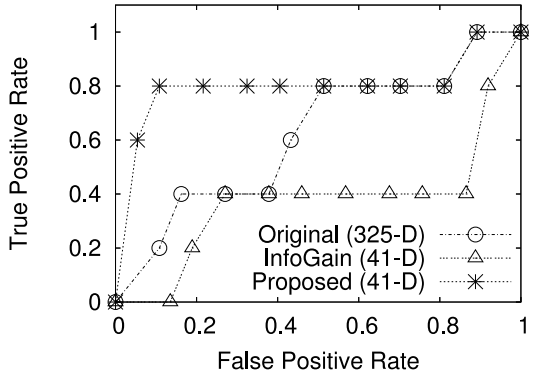


Figure. 5: Performance on Lung cancer data (using AVF method)

V. Conclusion and Future Work

A novel feature selection algorithm has been proposed here for effective detection of outliers in high dimensional categorical data. The proposed unsupervised algorithm is designed to evaluate the relevance and the redundancy of a feature independently. It first computes the entropy of a feature for assessing its relevance for outlier detection and then each relevant feature is evaluated for its redundancy with the features in the already selected subset. By characterizing the redundancy among the features through the mutual information computation, the proposed algorithm results in a selected subset of features with less redundancy. This algorithm has been experimentally evaluated on various benchmark categorical data sets employing two existing methods for outlier detection. As evident from the results obtained in this evaluation, the proposed algorithm has resulted in an efficient selection of features, leading to comparable outlier detection performance with that of the information gain-based supervised feature selection method.

Further work in this direction could be on employing more

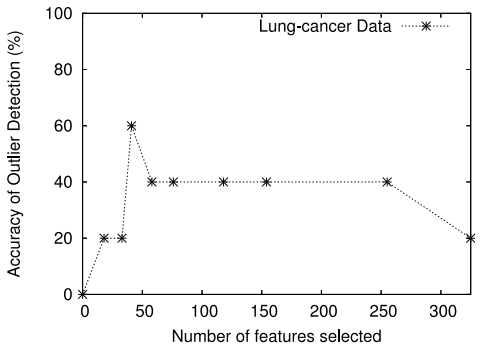


Figure. 6: Effect of number of features selected

useful measures of feature redundancy for outlier detection specific to a given application. As brought out in [28], the performance of the NMIFS method degrades in problems where group of features are relevant but not the individual features composing the group. Accordingly, the suitability of the hybrid filter/wrapper method proposed in [28] needs to be explored for dealing with the outlier detection problem. It is important to note that the algorithm proposed in this paper is only meant for use in connection with the outlier detection problem, but not for using as a generic feature selection method in other machine learning applications.

Acknowledgments

The authors would like to thank Director, CAIR for supporting this research work.

References

[1] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, pp. 85–126, 2004.

[2] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, 2009.

[3] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori, "Statistical outlier detection using direct density ratio estimation," *Knowledge and Information Systems*, vol. 26, no. 2, pp. 309–336, 2011.

[4] A. M. Bartkowiak, "Anomaly, novelty, one-class classification: A comprehensive introduction," *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 3, pp. 61–71, 2011.

[5] N. N. R. R. Suri, M. N. Murty, and G. Athithan, "Data mining techniques for outlier detection," in *Visual Analytics and Interactive Technologies: Data, Text and Web Mining Applications*, Q. Zhang, R. S. Segall, and M. Cao, Eds. New York, USA: IGI Global, 2011, ch. 2, pp. 22–38.

[6] M. Breunig, H. Kriegel, R. Ng, and J. Sander, "Lof: Identifying density-based local outliers," in *ACM SIGMOD ICMD*, Dallas, Texas, 2000, pp. 93–104.

[7] H. P. Kriegel, P. Kroger, E. Schubert, and A. Zimek, "Outlier detection in arbitrarily oriented subspaces," in *12th International Conference on Data Mining (ICDM)*. Brussels, Belgium: IEEE Computer Society, 2012, pp. 379–388.

[8] E. Muller, I. Assent, P. Iglesias, Y. Mulle, and K. Bohm, "Outlier ranking via subspace analysis in multiple views of the data," in *12th International Conference on Data Mining (ICDM)*. Brussels, Belgium: IEEE Computer Society, 2012, pp. 529–538.

[9] H. V. Nguyen, E. Muller, J. Vreeken, F. Keller, and K. Bohm, "CMI: An information-theoretic contrast measure for enhancing subspace cluster and outlier detection," in *SIAM International Conference on Data Mining (SDM)*, Asustin, Texas, 2013, pp. 1–9.

[10] N. Pham and R. Pagh, "A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data," in *18th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. Beijing, China: ACM, August 2012, pp. 877–885.

[11] A. Koufakou, E. Ortiz, and M. Georgiopoulos, "A scalable and efficient outlier detection strategy for categorical data," in *IEEE ICTAI*, Patras, Greece, October 2007, pp. 210–217.

[12] Z. He, X. Xu, and S. Deng, "A fast greedy algorithm for outlier mining," in *PAKDD*, Singapore, 2006, pp. 567–576.

[13] K. Das and J. Schneider, "Detecting anomalous records in categorical datasets," in *ACM KDD*, San Jose, California, 2007.

[14] S. Wu and S. Wang, "Information-theoretic outlier detection for large-scale categorical data," *IEEE Trans on Knowledge and Data Engineering (TKDE)*, vol. 25, no. 3, pp. 589–602, 2013.

[15] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection for discrete sequences: A survey," *IEEE Trans on Knowledge and Data Engineering (TKDE)*, vol. 24, no. 5, pp. 823–839, 2012.

[16] Q. Wu and S. Ma, "Detecting outliers in sliding window over categorical data streams," in *8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. IEEE Xplore, 2011, pp. 1663–1667.

[17] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in *Proc. of ACM SIGMOD International Conference on Management of Data*, Santa Barbara, USA, 2001.

[18] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[19] N. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special issue on learning from imabalanced data sets," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004.

[20] T. de Vries, S. Chawla, and M. E. Houle, "Finding local anomalies in very high dimensional space," in *IEEE ICDM*, 2010, pp. 128–137.

[21] H. V. Nguyen and V. Gopalakrishnan, "Feature extraction for outlier detection in high-dimensional spaces," in *JMLR WCP on Feature Selection in Data Mining*, vol. 10, 2010, pp. 66–75.

[22] M. Yamada, W. Jitkrittum, L. Sigal, E.P.Xing, and M. Sugiyama, "High-dimensional feature selection by feature-wise kernelized lasso," in *arXiv:1202.0515v2[stat.ML]*, Jun 2012.

[23] F. Keller, E. Muller, and K. Bohm, "Hics: High contrast subspaces for density-based outlier ranking," in *IEEE 28th International Conference on Data Engineering (ICDE)*. IEEE Xplore, 2012, pp. 1037–1048.

[24] T. M. Khoshgoftaar and K. Gao, "Feature selection with imbalanced data for software defect prediction," in *Proc. International Conference on Machine Learning and Applications*, 2009, pp. 235–240.

[25] M. Wasikowski and X. Chen, "Combating the small sample class imbalance problem using feature selection," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 22, no. 10, pp. 1388–1400, 2010.

[26] A. Lazarevic and V. Kumar, "Feature bagging for outlier detection," in *ACM KDD*, Chicago, USA, 2005, pp. 157–166.

[27] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[28] P. A. Estevez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Trans. on Neural Networks*, vol. 20, no. 2, pp. 189–201, 2009.

[29] T. Suzuki and M. Sugiyama, "Sufficient dimension reduction via squared-loss mutual information estimation," in *JMLR workshop and conference proceedings*, vol. 9, 2010, pp. 804–811.

[30] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in Neural Information Processing Systems 18*, Vancouver, Canada, December 2005.

[31] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Trans on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 301–312, 2002.

[32] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington DC, USA, July 2010, pp. 333–342.

[33] N. N. R. R. Suri, M. N. Murty, and G. Athithan, "Unsupervised feature selection for outlier detection in categorical data using mutual information," in *12th International Conference on Hybrid Intelligent Systems (HIS)*. Pune, India: IEEE Xplore, 2012, pp. 253–258.

[34] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans on Neural Networks*, vol. 5, pp. 537–550, 1994.

[35] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1997.

[36] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[37] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

Author Biographies

N N R Ranga Suri received his Bachelors and Masters Degrees in Computer Science and Engineering in 1998 and 2003 respectively. He has been working as a scientist with the Centre for Artificial Intelligence and Robotics (CAIR), Bangalore for the past 13 years. He has currently registered for PhD at the Indian Institute of Science (IISc), Bangalore. His present research work includes the development of Data Mining based methods for Information Security applications. He received the *Young Scientist* award from the Defence Research and Development Organisation (DRDO) for his contributions to various Data Mining oriented projects at CAIR.

M Narasimha Murty received his Ph.D. from the Indian Institute of Science, Bangalore, India in 1982. He is a professor in the Department of Computer Science and Automation at the Indian Institute of Science, Bangalore. He has guided 18 Ph.D. students in the areas of Pattern Recognition and Data Mining. He has published around 120 papers in various journals and conference proceedings in these areas. He worked on Indo-US projects and visited Michigan State University, East Lansing, USA and University of Dauphine, Paris. He is currently interested in Pattern Clustering.

G Athithan received his B.E (Hons) degree in electronics and communications engineering in 1981 from the Coimbatore Institute of Technology, Coimbatore, Tamilnadu. He received his Ph.D. degree in physics (of neural networks) in 1997 from the Indian Institute of Technology, Bombay. After completing the training at Bhabha Atomic Research Centre, Mumbai, he joined the Indira Gandhi Centre for Atomic Research (IGCAR), Kalpakkam, Tamilnadu in August 1982. He worked there for six years in computer graphics, computer aided design, and modelling of crystal structures. In October 1988 he joined the Advanced Numerical Research and Analysis Group (ANURAG), Hyderabad where he continued his work on computer graphics and visualisation besides taking up projects on parallel processing and neural networks. From June 2000 he was with the Centre for Artificial Intelligence and Robotics (CAIR) working in the field

of information security among others. Since Jan 2013 he is with the Scientific Analysis Group, Delhi. His current interests are information security management, computational intelligence, and network data mining and forecasting. He has published about twenty papers in archived journals and twenty five papers in national and international conferences.