Pattern Recognition Letters 32 (2011) 1054–1061

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Feature extraction for novelty detection as applied to fault detection in machinery

Jordan McBain, Markus Timusk *

School of Engineering, Laurentian University, Canada

ARTICLE INFO

Article history:
Received 29 September 2009
Available online 21 March 2011
Communicated by R.C. Guido

Keywords:
Novelty detection
One class classification
Feature selection
Feature reduction

ABSTRACT

Novelty detection is a pattern recognition technique used when there is one well characterized normal state and the abnormal (or novel) states are poorly described because of lack of data. Data deficiency of these states may arise due to cost and difficulty in measuring them – e.g. failed equipment states in equipment health monitoring. Normal pattern recognition techniques have a wide array of methods for reducing the number of features initially employed to characterize classes. These techniques are of limited use in novelty detection primarily because they are focused on representing the data accurately in a subspace rather than on finding a subspace where the classes can easily be discriminated, or they are optimized to distinguish between all classes rather than on focusing on distinguishing solely between normal and abnormal classes. The proposed methodology will enable feature extraction in unbalanced classification problems where the well-sampled normal data is expected to be orbited by the under-sampled fault data. The technique will be demonstrated to work well with artificial and actual machinery health data.

1. Introduction

In order to classify between states represented by groups of data points, one would normally seek as many features as possible to make up the components of a classification vector. Unfortunately, increasing the number of features results in an exponential increase in the number of sample points needed to characterize the problem's space; a result termed the *curse of dimensionality* (Duda et al., 2001). “The fundamental reason for the curse of dimensionality is that high-dimensional functions have the potential to be much more complicated than low-dimensional ones, and that those complications are harder to discern” (Duda et al., 2001). Many techniques have been developed in order to select the best subset of features to be processed by a pattern recognition method. Pattern recognition subject to data imbalance stands to benefit greatly by improvements in feature reduction techniques optimized for one-class classification problems.

The majority of classification algorithms assume relatively balanced data between classes. Classification problems that breach this tacit assumption are ill suited to such algorithms (Alejo et al., 2008). The learning objectives of these algorithms focus on classification accuracy of the training data set; under-sampled classes are poorly represented during mean-squared error like decision rules and the resultant classification rule becomes unduly biased (Sun et al., 2009). The objective of learning in this instance of pattern recognition is to provide higher identification rates of the under-sampled classes with a rate for the normal ones that is comparably stable (Zhu and Song, 2010). Real world problems include network intrusion detection, fraud detection, response modeling, etc. (Lee and Cho, 2006). The imbalance problem can be aggravated by factors such as small sample size, poor class separability, and within-class concepts (where class samples better reflect subcomponents of the class rather than the entire class itself) (Sun et al., 2009). Typical solutions to the problem include resampling some of the classes, adapting existing algorithms, boosting (a technique similar to employing a combination of classifiers), cost-sensitive learning or one-class learning.

Most pattern-classification techniques involve grouping representative feature vectors of different classes together and developing techniques to separate each class from the others. In the event, however, that the distribution of available data between classes is unbalanced, the ability of these techniques to distinguish between classes is limited. In machinery monitoring, this unbalance is particularly potent as data describing fault conditions is often nonexistent and acquiring it would require intentionally damaging equipment – a costly or completely unacceptable operational consideration. Novelty detection offers the solution by modeling the normal behaviour of the machine and triggering alarms when its operation falls outside of the normal model (a state presumed indicative of faults) (Markou and Singh, 2003).

While feature reduction, selection or extraction techniques employed in typical pattern recognition problems have been used with some success with novelty detection (Timusk et al., 2008), they are not optimized for this domain and there is room for much

* Corresponding author. Tel.: +1 705 675 1151x2243; fax: +1 705 675 4862.
E-mail address: mtimusk@laurentian.ca (M. Timusk).

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.01.019

improvement. One of the initial techniques employed was principal component analysis (PCA), which seeks to find some lower dimensional subspace where the data are best represented (Duda et al., 2001). Its failing was clear, in that it did not focus on finding a subspace in which data are best separated – a more suitable goal for pattern recognition; to address this concern, multiple discriminant analysis (MDA) was introduced. Its focus, however, is on finding a subspace in which *all* classes are optimally separated. This work will propose a method for novelty detection seeking a subspace where the normal and faulted classes are optimally separated; regrettably, the strongest form of the suggested technique is dependent on some data from faulted classes. The power of the approach will be demonstrated with artificial data sets and real machinery health monitoring data.

2. Feature reduction techniques

There are two primary thrusts in reducing a pattern classification problem's dimensionality: feature selection and extraction. Feature selection is the problem of choosing small subsets of features that are adequate to describe classes (Kira and Rendell, 1992). Feature extraction computes a small number of new features from the set of old features (Tax, 2001). Most of these techniques are not optimized for novelty detection; they are examined strictly to provide a background and understanding of the domain as a basis for understanding the new approach.

2.1. Search

Exhaustive feature search involves examining all subsets of feature combinations to find the one which maximizes some objective function (possibly the classification error achieved after having trained a classifier over that subset). This involves an exhaustive search over an exponential number of combinations that may not be practical depending on the definition of the objective function and the number of initial features. If computationally feasible, this method has been shown to find a better subset than most other methods available today (Timusk, 2006).

There are number of variants of search techniques employing heuristics that perform more optimally but are still computationally intensive; more on these methods can be found in Kira and Rendell (1992) and Timusk (2006).

2.2. Principal component analysis and variants

PCA seeks a subspace in which the data representation error is minimal. Duda et al. (2001) provide an excellent geometric motivation for this methodology that this review will repeat.

For a set of n vectors in d-dimensional space, we seek the equation of a hyper plane onto which the data may be projected with minimal representation error. The hyper plane is fixed at the data's mean, $\vec{m}$; the one point that can be demonstrated to represent the dataset optimally. The hyper plane's orientation is defined by the vector, $\vec{w}$. In the one dimensional case, the equation of the line is

$$\vec{x} = \vec{m} + a\vec{w} \tag{1.1}$$

Any point, $\bar{x}_k$, can be represented by this equation and the optimal set of coefficients a_k can be found by minimizing the squared error of

$$J(a_1 \ldots a_k, \vec{w}) = \sum_{k=1}^{n} \|(\vec{m} + a_k\vec{w}) - \vec{x}\,\|^2 \tag{1.2}$$

by minimizing it and solving for a_k. In order to determine the optimal direction, $\vec{w}$, the scatter matrix, S, ($n-1$ times the sample covariance matrix) is substituted it into $J(\vec{w})$, which then reduces to

$$J(\vec{w}) = -\vec{w}^t S \vec{w} + \sum_{k=1}^{n} \|x_k - m\|^2 \tag{1.3}$$

This expression can in turn be minimized by finding the minimum of $-\vec{w}^t S \vec{w}$; this is a well known problem in linear algebra and is solved by finding the eigenvectors, $\vec{e}$, that minimzes $J(\vec{w})$. From a statistic's perspective, these eigenvectors are the principal components of the hyper ellipsoidal distribution of the data's covariance matrix.

A feature subspace derived from this method is linear and may not well represent non-linear data and changing data. Kernel PCA and Dynamic PCA are recent additions to the literature designed to address these issues (Choi and Lee, 2004). Ultimately, however, these techniques are only suitable for representing data – not discriminating between them (Duda et al., 2001).

2.3. Multiple discriminant analysis

Discriminant analysis contrasts PCA in that it seeks to find efficient subspaces for discrimination rather than representation. In a two class classification problem with a set of d-dimensional points $x_1 \ldots x_n$, grouped into subsets D_1 and D_2, and projected onto some direction vector $\vec{w}$ to give

$$y = \vec{w}^t\,\vec{x} \tag{1.4}$$

that may be correspondingly grouped into subsets Y_1 and Y_2, one must find the direction vector $\vec{w}$ such that the distance between projected sample means, $\tilde{m}_1$ and $\tilde{m}_2$, is maximized. This distance must be rationalized against the relative size of each sample; instead, one maximizes the distance between means first rationalized for the total scatter of each class

$$J(\vec{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \vec{\tilde{s}}_2^2} \tag{1.5}$$

This expression can be shown to reduce to

$$J(\vec{w}) = \frac{\vec{w}^t S_{\mathrm{B}} \vec{w}}{\vec{w}^t S_{\mathrm{W}} \vec{w}} \tag{1.6}$$

where the numerator, containing the in-between class scatter matrix, S_{B}, can be derived by expanding the numerator of the previous objective function; and the denominator, containing the within-class scatter matrix S_{w}, can similarly be determined. The solution is described as analogous to that of the well known Rayleigh quotient and is given by Duda et al. (2001):

$$\vec{w} = \vec{S}_{\mathrm{W}}^{-1}(\vec{m}_1 - \vec{m}_2) \tag{1.7}$$

In the regular course, data are then projected onto the line defined by $\vec{w}$ and a novelty boundary is defined using a statistical threshold in the one dimensional subspace. It is not entirely practical to try and use the reduced data in the context of other non-statistical classification techniques such as the SVDD, SVM, or variants of neural networks. Interestingly enough, however, the approach has been extended to non-linear mappings in a technique termed kernel Fisher Discriminant Analysis (kFDA); the originators of the approach claim it works as well as some advanced support vector techniques (Mika et al., 1999).

Multiple discriminant analysis can also be extended into problems with n-classes. In which case, the objective is to maximize the spread between all classes in the projected space. This is an excellent extension of the technique but its objective function is too inclusive for novelty detection. In novelty detection, one seeks simply to distinguish between normal and abnormal states – not the normal class and all the substates of the abnormal state.