JOINT SUB-CLASSIFIERS ONE CLASS CLASSIFICATION MODEL FOR AVIAN INFLUENZA OUTBREAK DETECTION

Jie Zhang, Jie Lu, Guangquan Zhang

Centre for Quantum Computation & Intelligent Systems Faculty of Engineering and Information Technology, University of Technology, Sydney PO Box 123, Broadway, NSW 2007, Australia

{jie.zhang, jie.lu, guangquan.zhang@uts.edu.au}

H5N1 avian influenza outbreak detection is a significant issue for early warning of epidemics. This paper proposes domain knowledge-based joint one class classification model for avian influenza outbreak. Instead of focusing on manipulations of the one class classification models, we delve into the one class avian influenza data-set, divide it into sub-classes by domain knowledge, train the sub-class classifiers and unify the result of each classifier. The proposed joint method solves the one class classification and feature selection problems together. The experiment results demonstrate that the proposed joint model definitely outperforms the normal one class classification model on the animal avian influenza data-set.

Keywords: One class classification; avian influenza; outbreak detection; joint model

1. Introduction

H5N1 avian influenza outbreak detection is a significant task with a big challenge, because there are a number of uncertain factors associated with the outbreaks [1] . The Asian lineage, highly pathogenic avian influenza (HPAI) virus sub-type H5N1 was first identified in Hong Kong in 1996 [2], and worldwide outbreaks increased dramatically since 2003. Now it has spread to more than 60 countries from Asian, to Europe and Africa [3]. This epidemic has infected poultry and caused the culling of millions of birds resulting in a loss of billions in the poultry trade, and significant loss of human life. According to WHO [4], 502 humans have been infected and among them 302 of them have died. This virus can mutate according to its host and adapt to different environments [5, 6]. Normally, the virus transmits from birds to mammals, but so far, it seems it cannot be transmitted between mammals [5]. Water birds are believed to be the viral reservoir of influenza A viruses [7] but the transmitting mechanism is still unclear. Wild birds and human activities have broadened the channels for the virus to spread; for example, poultry farm, bird trade and wild birds' migration [5, 8, 9] are all possible channels to spread the virus. In south Asia, free-grazing duck are also believed to be a major reason for virus transmission [10]. Scientists are still struggling to discover the reasons for transmission.

However, if we take outbreak events as our target observation and the normal status without outbreaks as the outliers, then detection turns into a typical one class classification (OCC) issue. The item "one class classifier" was first proposed by Moya [11] and one class classification (OCC) has been studied in the area of the novelty detection [12], outlier detection [13] in signal processing and pattern recognition applications. At the outset, the focus of OCC is to identify novelties and get rid of them before processing

normal signals. This focus subtly changes to analyzing the target class in the research area of data mining [14], and OCC method has been developed and greatly advanced by Tax and Duin [15] by their idea of one class data description. OCC in data mining depicts the only labeled target class by a suitable model and detects the new case if it is in the boundary of the target class or if it is out of the boundary as an outlier.

The situation of OCC is very common in real world. These one label tasks can often be encountered in the real world, for example, nuclear plant failure, a medical disease case, and identifying a type of web pages [16, 17]. OCC has been widely used in many areas, such as cyber-intrusion detection, medical diagnosis, image processing, fraud detection in financial industry [17], and defect detection in the fabric industry [18]. Other applications include land cover classification [19], environmental monitoring [20] , document retrieval and classification [21, 22], vague stream data analysis [23] and the most promising application domain, which we will discuss next, is the medical and biological area.

OCC methodology is especially suitable for medical and biological domain applications. Duin has mentions that OCC has been used in disease detection [24]. In many medical diagnoses, the doctor only keeps the disease case data which can be used as the labeled target class, and all other diseases and healthy cases are taken as the outliers. In the area of epidemic disease for instance, avian influenza, animal cases are only reported if there are epidemic outbreaks in a poultry farm or a number of dead wild birds are identified. This is similar in OCC applications applied in gene science [25-28] where only target gene samples are available. The situation in these areas is approximately the same and researchers only focus on limited target samples, while outliers have a large population, such as healthy populations versus target disease subjects, or RNA of all other animals versus the target RNA gene pattern.

Though OCC methods have many applications, there are limitations due to the nature of only one labeled class. Many researchers choose two-class or multi-class data-sets in their experiments [25, 26, 29, 30] to analyze the results. Other researchers artificially generate outliers for evaluation [27, 31]. In real world examples, we can only obtain the target class [28] such as avian influenza outbreaks. The outbreaks are a typical OCC issue with only one label, whilst other issues need to be addressed in depicting avian influenza animal outbreaks: 1) All the features leading to the outbreak are unclear because the H5N1 virus has mutated to adapt to the environment and the virus distribution channels are complicated [5]. 2) The outbreak happens incidentally. This makes it hard to tell exactly the differences between the outbreak and the non-outbreak cases. The above two reasons make outbreak detection very difficult. First, the feature space is uncertain and hard to evaluate. The factors causing an outbreak are not clearly identified and refining of factors is difficult because we have only the target labels. Secondly, the possibilities of outbreaks decrease the necessity and the effectiveness of generating artificial outliers.

Under this circumstance, we proposed an ensemble classifiers approach, joint sub-classifier one class classification (JSC-OCC) method, to overcome these difficulties. The

ensemble classifiers have long been studied [32] and have been verified to improve the performance of the classification [33]. Contemporary research for ensemble classifier is represented by bagging [34] and boosting [35] methods, and the extension of the two methods. These methods make full use of the original dataset by a different method of sampling to make the classification results more stable and accurate. However, this research has some limitations: Firstly, these methods have limitations of improving the individual base classifiers by the learning domain knowledge [36]. In a real world dataset, the domain knowledge will significantly affect the classifying result, which is why we apply domain knowledge into the classification; secondly, many applications only focus on multi-class classification. There are many research applications on combination of classifiers [37, 38], but these applications mostly address multi-class classifications. In the real world, there are many one class classification problems which focus on the target class identification and are different to multi-class classification problems, which treat classes equally. Instead of seeking assistance from the second class in this research, we delve into the target cases by classifying the outbreak cases into sub-classes, training the classifiers separately and integrating the separated models. We apply a supervised feature selection method to the sub-class and achieve better results with only half of the features selected.

The paper is organized as follows. Section 2 describes the OCC methods, rated OCC research work and reviews the limitations of the previous methods. Section 3 presents the motivation and the contents of the JSC-OCC method. Section 4 illustrates the method on the avian influenza animal outbreak data-set and the results show that the JSC-OCC method out-performs the normally applied OCC method. Section 5 concludes the paper.

2. OCC METHODS AND RELATED WORK

There are many OCC methods which have been applied in many real world applications. We describe the concept of the OCC method and the related research results.

2.1. *OCC methods*

The basic idea of OCC is described by two essential elements. The first is the distance, or the possibility of a new case for the target class. The second element is the threshold on this distance or possibility. Whether or not the new case belongs to the target class can be defined by the distance, less a threshold:

$$f(\mathbf{y}) = I(d(\mathbf{y}) < \theta_d) \tag{1}$$

or the possibility is larger than the possibility threshold:

$$f(\mathbf{y}) = I(p(\mathbf{y}) > \theta_p) \tag{2}$$

where $I(.)$ is the indictor function, y is the new case, $d(\mathbf{y})$ and $p(\mathbf{y})$ is the distance and the possibility on the target class. θ_d and θ_p is the distance and possibility threshold.

Different OCC method have different definitions of $d(\mathbf{y})$ or $p(\mathbf{y})$ and the evaluation of the OCC method can compare the hypercube or the hypersphere volumes depicted by $d(\mathbf{y})$ or $p(\mathbf{y})$. The threshold value is a trade off between whether a new coming case is a target class or an outlier.

OCC methods can be divided into three main categories: the density method, the boundary method and the reconstruction method [39]. The density method applies statistical distribution to depict the target class probability density, and uses a threshold to distinguish targets from outliers. The density method examples are the Gaussian data description (GaussianDD) model, mixture Gaussian data description (MOGDD) and Parzen data description (ParzenDD) model. The boundary method draws a boundary including the targets with the minimum volume, such as support vector data description (SVDD) model and k-nearest neighbor data description (KNNDD) model. The reconstruction method reconstructs previous space by the prototype model or data compress model and identify targets and outlier after reconstruction, e.g. k-means data description (KmeansDD) model, principal component analysis data description (PCADD) model, or the self-organizing map data description (SOMDD) model. In a real world application, different models will have different performances according to the implementation [39]. The minimum spanning tree data description (MSTDD) method is a new promising method which out-performs many other previous one class data description methods by defining the target boundary of the minimum volume around the spanning tree [24].

GaussianDD is a basic OCC density method, which apply Gaussian distribution to describe the one class case according to the Central Limit Theorem. The possibility of *d*-dimensional targets $\mathbf{x}$ is given by:

$$p(\mathbf{x},\mathbf{u},\mathbf{\Sigma}) = \frac{1}{(2\pi)^{d/2}\left|\mathbf{\Sigma}\right|^{1/2}}\exp\{-\frac{1}{2}(\mathbf{x}-\mathbf{u})^{T}\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{u})\} \tag{3}$$

where $\mathbf{u}$ is the mean value and $\mathbf{\Sigma}$ is the co-variance matrix. The target class should be a strict unimodal and convex density distribution. The main calculation cost is the inversion of the matrix $\mathbf{\Sigma}$. The other density models are the extension of the GaussianDD model.

SVDD model is a typical boundary OCC method which is different from one class support vector machine (SVM). One class SVM turns an OCC method into a two class classification method by defining the origin as a second class [17]. But the SVDD model draws a minimum volume hypersphere to contain most of, or all of, the targets. SVDD is an optimization problem with the object as:

$$\min. f(R,\mathbf{a},\xi) = R^2 + C\sum_{i=1}^{N}\xi_i \ , \tag{4}$$

$$s.t. \quad (\mathbf{x}_i - \mathbf{a})^T (\mathbf{x}_i - \mathbf{a}) \le R^2 + \xi_i \quad i = 1,..., N, \xi_i > 0 \tag{5}$$

where $\mathbf{a}$ is the center and R is the radius of the hypersphere, ξ_i is the slack variable and C is the variable to describe the trade off between the sphere volume and the number of the target objects rejected. After applying Lagrange multipliers to the problem we will have the dual problem:

$$L = \sum_{i=1}^{N} \alpha_i (\mathbf{x}_i . \mathbf{x}_i) - \sum_{i,j=1}^{N} \alpha_i \alpha_j (\mathbf{x}_i . \mathbf{x}_j) \tag{6}$$

where $0 \le \alpha_i \le C, i = 1,..., N, \sum_{i=1}^{N} \alpha_i = 1, \sum_{i=1}^{N} \alpha_i \mathbf{x}_i = \mathbf{a}$

We can predict if a new case is accepted or not by:

$$\begin{aligned} \|\mathbf{y} - \mathbf{a}\|^2 &= (\mathbf{y} - \mathbf{a})^T (\mathbf{y} - \mathbf{a}) \\ &= (\mathbf{y} - \textstyle\sum_i \alpha_i \mathbf{x}_i)^T (\mathbf{y} - \sum_i \alpha_i \mathbf{x}_i) \\ &= (\mathbf{y}.\mathbf{y}) - 2\textstyle\sum_i \alpha_i (\mathbf{y}.\mathbf{x}_i) + \sum_{i,j} \alpha_i \alpha_j (\mathbf{x}_i . \mathbf{x}_j) \le R^2 \end{aligned} \tag{7}$$

If we substitute all the inner products $(\mathbf{x}_i . \mathbf{x}_j)$ with a kernel function:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i).\phi(\mathbf{x}_j) \tag{8}$$

Then the problem can be mapped into an inner product space.

K_meansDD method is a very simple reconstruction method. It applies prototype vectors $\boldsymbol{\mu}_k$ to minimize the error:

$$\varepsilon = \sum_i (\min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2) \tag{9}$$

Where $\boldsymbol{\mu}_k$ is the vector of k-means center, $\mathbf{x}_i$ is the target case vector. Different reconstruction methods have different error measuring methods. More details of SVDD and other OCC methods are described by Tax [39].

Many new OCC methods and improvements have been discovered. These methods can be mainly categorized as localization and combinations. Localization tunes the parameters according to the local properties which differ from the global ones. The localization method has been applied to KNNDD [17] and Gupta [29]; it combines local and global searches to improve the one-class information ball (IC-IB) method. The combination method combines different models to obtain better results and has been

explored in OCC modeling. Combining density estimator with two class probability estimator has been proposed [30]. The normal ensemble methods are applied in gene science [26] and the combination of the OCC model with different data-sets has also been studied [27]. Statistical learning and case-based reasoning combination methods have also been proposed to integrate the similarity measure and Bayesian statistical information to identify novelties [40]. Only a few research methods deal with the genuine one class issue [30], where it is impossible or improbable to obtain an outlier. This situation will make many proposed methods unsuitable, including information from the outliers.

2.2. *Feature selection issue in OCC*

Feature selection chooses high variance features and removes low variance ones, but target class label provides no information at all [31]. This means only unsupervised feature selection methods can be applied under this circumstance. Villalba [41] evaluated four feature selection algorithms and concluded that the Q-α algorithm and locality preserving projections (LPP) have better performance. Generally, unsupervised methods cannot compare with supervised ones. Most unsupervised dimension reduction methods just compress the original feature space into a smaller dimensions, which hardly explains the meaning of each dimension, for example LPP and PCA method [42].The most promising supervised feature selection method is mutual information feature selection, for instance the minimum-redundancy maximum-relevancy (mRMR) method [43], which makes full use of mutual information between the features and class labels to select the feature group with most variation.

The above review indicates that we can investigate limitations of current OCC methods in dealing with real world examples. Avian influenza outbreak event detection is this type of issue. Firstly, a genuine OCC problem can only obtain the target label, which means that we cannot obtain all the other class data without outbreak events. If we can provide some definite outliers for the OCC models, the classifier can tune [31] and the result will be significantly improved. This means that many OCC techniques which tune on the second class samples are not suitable. Secondly, the supervised features selection should not be applied because the outbreak factors are unclear and we need to select features. Therefore, we can only apply the unsupervised feature selection method. We next present our own approach for dealing with this real world OCC problem.

3. Joint Sub-Classifier OCC Method

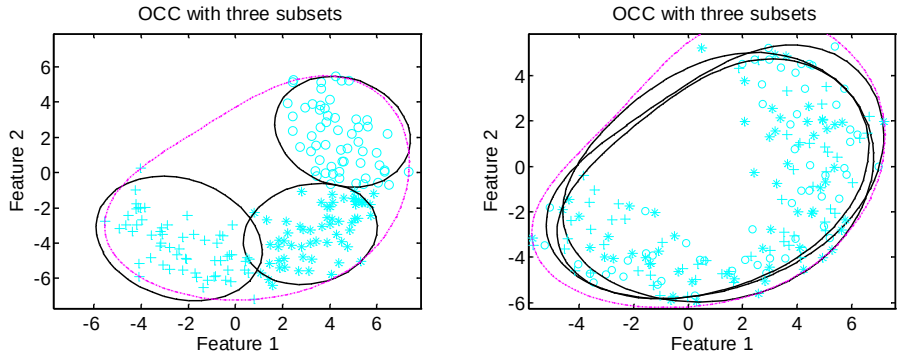
We describe the motivation and the detailed processes of our JSC-OCC method step-by-step.

3.1. *The motivation of proposed method*

The OCC method is designed for resolving the classification problem when the training data-set only has one label. If there is only one class labeled as target, e.g. the avian influenza outbreak events, we can improve the OCC method with the following method.

We divide the outbreak events into sub-categories to discover whether the new sub-class of the outbreak events can help improve the OCC models to detect if a new coming case is likely to be an outbreak or not. On one hand, if the sub-classes can be grouped closely together and at same time apparently appear separate to each other, then we can apply the OCC method for each sub-class, then combine every OCC sub-classifier to obtain a better result. On the other hand, if the sub-groups cannot be separated from each other, then we should select the most suitable features to help classify the sub-classes.

The basic idea is shown very clearly in Fig. 1 by a Two-D SVDD method. We provide a well separated three sub-class example in Fig. 1 (a). The three black line circles are the SVDD boundaries for the sub-classes. The magenta dashed line is the SVDD boundary for the labeled one class. In this context, the joint three circle boundaries are better than the one big circle because they have less volume. Therefore, if we can find a well separated sub-class and the features which can help separate the sub-class, then we can improve the detection effectively. In Fig. 1(b), we cannot observe that the combination of sub-class circles improved the precision of the OCC method, or the combination cannot reduce the volume of the OCC boundary. This means that the three sub-groups on this feature space cannot improve the classification effectiveness.



a) three well separated sub-classes and features b) three poorly separated sub-classes and features

Fig 1. One class SVDD and three separated sub-classes SVDD

Therefore, we must address two issues: first, we delve into one class and divide it into well separated sub-classes; second, we select good features to improve the separated effect. In the medical domain, epidemic domain and biological domain, the first task is naturally completed by the domain expert. In the medical domain, the doctor will separate samples of an illness into slight, and severely ill groups, male and female groups, or young and old categories. In the epidemic disease domain such as avian influenza, the sub-class can be wild bird affected events and poultry farm events. This means the first task is resolved by domain knowledge. The second task is also easy to accomplish because we can select the features according to the sub-class labels. The

previous unsupervised feature selection task becomes a supervised feature selection and we can easily resolve the two problems at the same time.

3.2. *Processes of the JSC-OCC method*

The combining method concludes the following 5 steps:

Step 1. Divide all the one class cases into sub-classes by domain knowledge;

Even though we have only one class label, we can divide the one class into sub-classes by applying domain knowledge. The effective sub-class needs to be carefully considered. In avian flu epidemic domain, there are many ways to group the outbreak events such as the outbreak seasons, the locations and so on. Nevertheless, if we consider the transmission of the virus and the mobility of the population, we can divide the affected population upon different sub-groups as wild species, backyard free range poultry and farm poultry. The wild birds have the maximum mobility, the backyard birds have limited mobility and the farm poultry has the less mobility. We don't need to divide the farm poultry as broiler chickens, layer chickens, or turkey and so on, because we consider that the transmission ways to the farm maybe similar, e.g. mainly by human activities.

The dividing task is fully depended on the domain knowledge and the research objective. If we have enough details of the domain knowledge, then the sub-classes will contain the more variation information. If the objective is different, the classification standard will be different. For example, if we need to undertake research into the vaccine effect of poultry, we must divide farm or free-range poultry into vaccinated and unvaccinated groups. If we only consider the transmission methods of the virus, we only need to divide the birds into the previous mentioned three sub-groups.

Step 2: Select the most variation features according to the sub-classes;

Here, we select features which help enforce the classification effect. In medical and epidemical areas, it is normal that there will be many factors associated with the disease case and is important to group the factors and select the suitable ones. In the normal OCC method, we apply only the unsupervised feature selection method, but here we can select features by either unsupervised or supervised methods because we have sub-class labels which allow selection of features based on the sub-class labels. This provides a lot more choice than previous OCC models, for example, we can apply mRMR method to select the features.

Step 3: Train the OCC classifier on each sub-class;

We apply each sub-class cases as training data-set to train sub-OCC models. We can either use the same OCC classifier on different sub-class data, or we can use different OCC classifiers on different sub-class data. We then obtain several sub one class classifiers.

Step 4: Combine three sub-classifiers to union into a joint OCC model.

We combine the sub-classifiers' results to obtain a joint result. We choose to join the results of the sub-classifiers by logical 'or' operator. Suppose we have n classifiers, for each new coming case x , the final result can be obtain by:

$$I(x) = \bigcup_{i=1}^{n} C_i(x) \tag{10}$$

C_i is the *i*th classifier, $C_i(x)$ is true if x is detected as a target class and $\boldsymbol{I}$ is the output of joining the output results. If one of the classifiers classifies the input case as target, the final output is the target class. The whole process is shown in Fig. 2

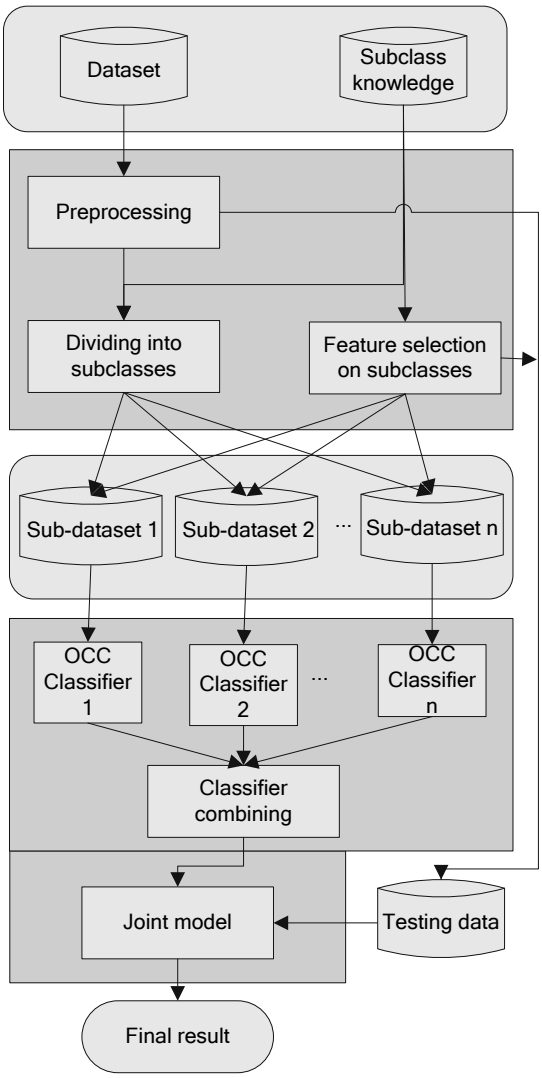


Fig 2 The JSC-OCC method processes

4. Experiment and Result Analysis

The experiment is conducted on the avian influenza data-set collected from the internet. Some features are transformed by GIS software before input to the model. We apply Matlab platform and DD_Tools [44] to perform the experiments.

4.1. *Data source*

One section of avian influenza animal outbreak data was obtained from reports on the website: http://www.oie.int/downld/AVIAN%20INFLUENZA/A_AI-Asia.htm; the other

data was obtained from Nature News reporter Dr Declan Butler (http://www.nature.com/news/author/Declan+Butler/index.html). Each record contains the outbreak time, outbreak location, infected population, location type, and so on. The data-set has records dated from 2003 to 2009. If the outbreak location is a farm, the dead bird numbers or infected numbers on the farm do not affect other farms. For processing purposes, we count this farm event as only one event. Dead wild bird events are counted as one event even if only one dead bird is identified. The other feature data is collected from the internet, such as poultry density, poultry and wild bird trade, population density, from http://www.fao.org ; other geographic data is collected from http://eros.usgs.gov/, http://www.ngdc.noaa.gov, etc. These data will first be pre-processed by GIS software, and then processed to obtain the feature information of an outbreak event by Matlab. The basic location data from OIE reports has errors, for example the wrong information of longitude and latitude, which indicates a location in the sea and we cannot obtain correct information from them, so record this as missing data. After removing the missing data, we have 5,600 cases. Each case has 24 features plus the affected bird type information. All these features are shown in Table 1.

Table 1 All the selected features

No	Features	Descriptions
1	'latitude'	Latitude
2	'logitude'	Longitude
3	'clim_tmp'	Temperature
4	'clim_wet'	Wet days
5	'clim_tmx'	Daily Maximum Temperature
6	'clim_tmn'	Daily Minimum Temperature
7	'clim_frs'	Ground frost frequency
8	'clim_vap'	Water vapour
9	'clim_dtr'	Diurnal temperature range
10	'clim_cld'	Cloud cover
11	'pop_den'	Population density
12	'pty_den'	Poultry density
13	'rail_den'	Rails line density
14	'mrail_den'	Main rails line density
15	'road_den'	Road line density
16	'water_line'	River line density
17	'hydro_den'	Main hydro point density
18	'pmeat_trd'	Poultry meat import volume
19	'bird_trd'	Live bird import volume
20	'elevation'	Elevation
21	'bat_cover'	BAT Land cover type
22	'month'	Month

23	'year'	Year
24	'migr_rsk'	Near how many migration routes
25	'bird_type'	Affected bird population type

These 25 features are possible features associated with the avian influenza outbreak. We select only 25, classified as climate factors, geographic factors, poultry density factor, transportation factors, water factor, bird trade and wild bird migratory information. For example, the bird migratory risk is defined by the following processes: there are eight main bird migratory routes around the world – the 'migr_rsk' features are calculated by how near the outbreak location is to how many the migration routes. If the minimum distance to a bird migratory route is less than 300 km, we count the 'migr_rsk' attribute once. The transportation level is measured by the line density per km^2 of railroad length and road length. The indicator of poultry meat trade is calculated from the quantities of tons of imported poultry meat divided by the country total area, and the same is done with the live bird trade.

4.2. *Experiment result and analysis*

Here, we follow the steps of the proposed JSC_OCC method and conduct the experiment on an avian influenza data-set.

Step 1. Divide all the one class cases into sub-classes by domain knowledge;

In avian flu outbreak events, the affected birds can be divided into three sub-classes as wild birds, backyard birds and farm poultry. We divide into these three sub-class taking into consideration the mobility of the populations. Wild birds have the greatest mobility, farm poultry has the lowest mobility, and backyard birds have limited mobility. The movements of birds should be connected with the transmission of the viruses. The 'bird_type' is chosen as the sub-class label and the remaining 24 will be applied by the OCC method. The whole data-set can be divided into 1,086 wild bird affected cases, 1,169 backyard poultry affected cases, and 3,345 farm poultry bird affected cases.

Step 2: Select the most variation features according to the sub-classes;

This step applies the mRMR feature selection method and we only choose 12 features, or half, of the total features to illustrate our approach. The 12 selected features are "'elevation', 'rail_den', 'bird_trd', 'pmeat_trd', 'clim_wet', 'road_den', 'mrail_den', 'year', 'clim_tmp', 'migr_rsk', 'logitude', 'month'".

Step 3: Train the OCC classifier on each sub-class;

The training data-set randomly selects 80 percent of each sub-class case. We also merge the three training data-sets in to a fourth data-set as a comparison. We then have four trained OCC models. The OCC models applied are GaussianDD model, MOGDD model,

ParzenDD model, SVDD model, 1NNDD model, KNNDD model, KmeansDD model, PCADD model and SOMDD model. We choose the rejected threshold as 0.1 and apply the default parameters of DD_tools.

Step 4: Combine three sub-classifiers to union into a joint OCC model.

We combine the trained three sub-classifiers by the results as (1). Finally, the test data is the same as the 20 percent of remaining data. The results of the experiments are shown in Table 2 and Table 3. Table 2 is the experiment with all the features and Table 3 illustrates the result with only 12 selected features.

Table 2 Experiment with all features

	training data				
	Wild bird (Tdata1)	***Backyard poultry (Tdata2)***	***Farm poultry (Tdata3)***	***Tdata4 =(Tdata 1+Tdat a2 +Tdata 3)***	
Correct classified rate of model on same test data	***Sub-OCC1 correct rate***	***Sub-OCC2 correct rate***	***Sub-OCC3 correct rate***	***OCC collect rate***	***JSC-OCC collect rate***
Gaussian	0.1891	0.4674	0.6976	0.9037	0.9304
Parzen	0.1436	0.1499	0.3702	0.5674	0.5932
Kmeans (k=5)	0.2417	0.5843	0.7208	0.8912	0.9322
knn	0.2971	0.4451	0.8073	0.8921	0.934
som	0.2614	0.4442	0.8127	0.9037	0.9349
nn	0.7538	0.9233	0.9269	0.9197	0.9973
mog(5)	0.1855	0.397	0.6789	0.8912	0.9144
Svdd (σ=100)	0.1508	0.6111	0.4906	0.6441	0.802
pca	0.215	0.5343	0.7056	0.8974	0.942
mst	0.5566	0.4853	0.1124	0.0321	0.8037

Table 3 Experiments results with only 12 features

	Training data				
	Wild bird (Tdata1)	***Backyard poultry (Tdata2)***	***Farm poultry (Tdata3)***	***Tdata4 =(Tdata 1+Tdat a2 +Tdata 3)***	
Correct classified rate of model on same test data	***Sub-OCC1 correct rate***	***Sub-OCC2 correct rate***	***Sub-OCC3 correct rate***	***OCC collect rate***	***JSC-OCC collect rate***
Gaussian	0.2087	0.5192	0.6878	0.9019	0.9438
Parzen	0.1561	0.1945	0.4648	0.6842	0.7047
Kmeans (k=5)	0.2774	0.5165	0.7145	0.9153	0.9402
knn	0.3363	0.3854	0.7966	0.901	0.9313
som	0.3301	0.4665	0.694	0.8965	0.9295
nn	0.7948	0.9135	0.9072	0.9224	0.9955
mog(5)	0.1847	0.4282	0.6967	0.8992	0.9242
Svdd (σ=100)	0.1641	0.521	0.4755	0.6798	0.7583
pca	0.2257	0.4987	0.7047	0.901	0.9634
mst	0.5709	0.5085	0.1133	0.0375	0.7895

The evaluation standard is obvious for comparing the correct classification rate to the target class, the true positive (TP) rate. In Table 2 and Table 3 the correct classification rates are rates on the whole testing data. We observe that the JSC-OCC method can outperform the normally applied OCC models and we conclude from the two tables that, whether we apply feature selection first or not, the JSC-OCC model will outperform the normal OCC model on the outbreak data-set. Though the sub-OCC models on the testing data have a low TP rate, the JSC-OCC model performances improve significantly, as we expected.

We also compare Table 3 to Table 2 with the TP rate of applying all the features, and only half of the features. From the data we observe that five out of ten combined models with the selected 12 features in the experiments outperform the combined models with all 24 features. These five combined JSC-OCC models are the GaussianDD, ParzenDD, K-meansDD, MogDD and PCADD models. But the results of JSC-OCC with only the 12 features data-set have more accuracy than the results of the normal OCC method on the 24 data-set. Five models gain better performances with selected features and this means the feature selections on the sub-class will not decrease the FP rate of JSC-OCC models. We haven't tuned the parameters of the SVDD model, so it has the lower accuracy. The 1NNDD sub-classifiers have very high accurate rate which means they are not sensitive

with the groups divided. But the JSC-OCC result based on 1NNDD still has better performance. We also notice that MSTDD performed poorly, which is said to outperform many other models. We explain this by pointing out that MSTDD is a more strict OCC method with the minimum volume boundary, which means it can only perform best when the outliers are identified clearly. In the real world one class problem, you really cannot provide a clear description of the outlier. So the strict MSTDD will perform poorly compared with other OCC models.

5. Discussion

The JSC-OCC method will improve the accuracy of the classification provided we have the domain knowledge to divide the whole population into sub-groups. Sometimes it is very difficult, or even impossible, to obtain the knowledge to divide the sub-classes. Under these circumstances, we can still classify the data-set by clustering. Whether or not the unsupervised sub-classes improve the accuracy of OCC models is an interesting problem for study. Clustering method is the most common unsupervised classification method, and we classify the outbreak cases into three sub-classes by k-means and the spectral clustering method separately, and conduct the JSC-OCC method on these two sub-classes. We do the experiments both on the 'all features' and selected' 12 features' data-sets and the results are listed in Table 4 and Table 5.

Table 4 K-means: Three sub-groups with all features:

	24 features			**12 features**		
Correct classified rate	***OCC***	***JSC-OCC***	***improved or not***	***OCC***	***JSC-OCC***	***Impoved or not***
Gaussian	0.8839	0.9062	TRUE	0.8786	0.8902	TRUE
Parzen	0.5839	0.5893	TRUE	0.6786	0.6759	FALSE
Kmeans (k=5)	0.8848	0.8902	TRUE	0.8768	0.8982	TRUE
knn	0.883	0.9116	TRUE	0.9054	0.9116	TRUE
som	0.8848	0.8911	TRUE	0.8938	0.8777	FALSE
nn	0.9304	0.9982	TRUE	0.933	1	TRUE
mog(5)	0.8911	0.8777	FALSE	0.8938	0.8759	FALSE
Svdd (σ=100)	0.6232	0.5839	FALSE	0.6357	0.7018	TRUE
pca	0.9062	0.9429	TRUE	0.8964	0.9509	TRUE
mst	0.0455	0.9982	TRUE	0.0375	0.9991	TRUE

Table 5 Spectral clustering: Three sub-groups:

Correct classified rate	**24 features**			**12 features**		
	OCC	***JSC-OCC***	***improved or not***	***OCC***	***JSC-OCC***	***Imporved or not***
Gaussian	0.8858	0.8956	TRUE	0.8787	0.9001	TRUE
Parzen	0.6021	0.6004	FALSE	0.6459	0.661	TRUE
Kmeans (k=5)	0.9019	0.9197	TRUE	0.8831	0.9135	TRUE
knn	0.9019	0.9019	FALSE	0.9019	0.9072	TRUE
som	0.9019	0.8956	FALSE	0.8938	0.8992	TRUE
nn	0.9411	0.9946	TRUE	0.9277	0.9938	TRUE
mog(5)	0.8956	0.8876	FALSE	0.8912	0.8876	FALSE
Svdd (σ=100)	0.6343	0.7145	TRUE	0.7074	0.6467	FALSE
pca	0.8983	0.9081	TRUE	0.8965	0.9108	TRUE
mst	0.0419	1	TRUE	0.041	1	TRUE

Table 4 lists the comparisons between results of JSC-OCC and OCC on the K-means three sub-groups and Table 5 lists the comparisons between results of JSC-OCC and OCC on the three spectral clusters. The experiments have been conducted with ten OCC models on both 'all features' and '12 features' data-sets. From Table 4, we observe that eight OCC models employing the JSC-OCC method perform well on the 24 features data-set and seven OCC models employing the JSC-OCC method perform better on the 12 features data-set. From Table 5 we also observe similar results with six out of ten JSC-OCC models performing better on 24 features data-set and eight out of 10 JSC-OCC models performing better on 12 features data-set. We cannot reach a conclusion that unsupervised sub-classes improve the accuracy of the JSC-OCC method because not all JSC-OCC methods provide better results than the normal OCC method.

However, the above experiments are conducted on the three sub-groups data-set. In fact, we can have different number of sub-groups, so we conduct the experiments on two to twenty sub-groups on the 24 features data-set to compare the performances between JSC-OCC method and normal OCC method. This time we only apply six OCC models as GaussianDD, MoGDD, ParzenDD, SOMDD, K-meansDD and KNNDD models. We still applied K-means and spectral clustering methods to cluster the data-sets. The results are shown in Figure 3 and Figure 4.

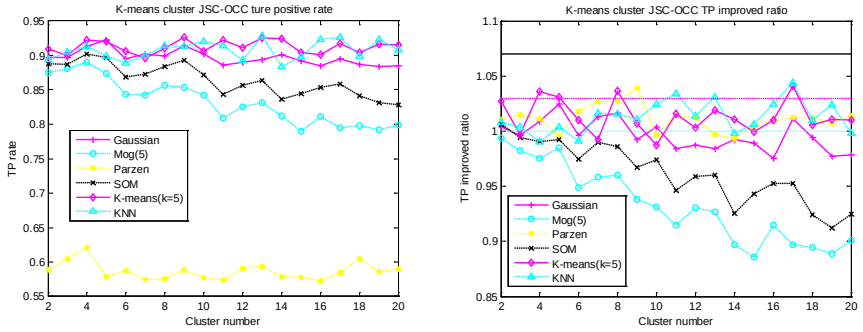


Figure 3 (a) JSC-OCC FP rate on different K-means sub-groups. (b) JSC-OCC FP rate improved ratio on different K-means sub-groups

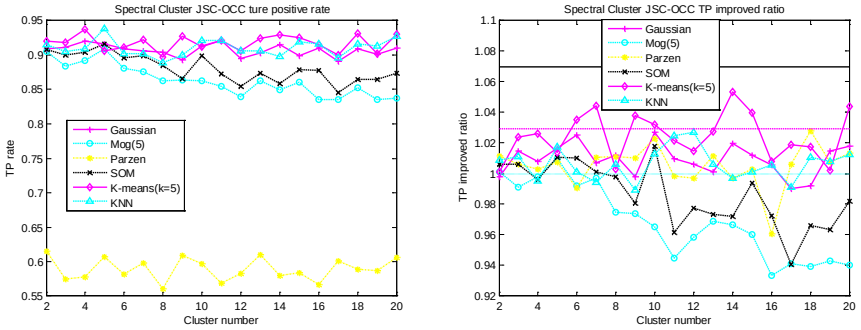


Figure 4 (a) JSC-OCC FP rate on different spectral clustering sub-groups. (b) JSC-OCC FP rate improved ratio on different spectral clustering sub-groups

The TP rates and TP rate improved ratio of JSC-OCC on different number of K-means sub-groups are shown in Figure 3 and similar results on spectral cluster sub-groups are shown in Figure 4. The TP rates of JSC-OCC are illustrated in Figure 3 (a) and Figure 4 (a). The TP rate improved ratio is calculated by the TP rate of JSC-OCC method divided by the TP rate of OCC method and is listed in Figure 3 (b) and Figure 4 (b). We find that TP rates of JSC-OCC method fluctuates with the variations of cluster number. There are no obvious trends with almost all the JSC-OCC models except TP rates of SOMDD and MoGDD models show a slightly decreased trend. The results are similar for both TP rate on K-means sub-groups and for TP rate on Spectral cluster sub-groups. We conclude that the unsupervised cluster number has no obvious effect on the TP rate of the JSC-OCC method and some OCC models performances degenerate. We also found that most of the TP rate improved ratios are above one and this situation is clearer in Figure 4 (b) with the TP rate improved ratios on spectral cluster sub-groups. This phenomenon indicates that TP results on most of the sub-groups of the data-set can be improved and it is most obvious with spectral clustering sub-groups.

The last observation from Figure 3 and Figure 4 is that the TP improved ratio is limited in comparison to the supervised sub-groups. The average TP improved ratio of supervised, or knowledge based, JSC-OCC method is 1.0699 and the minimum TP improved ration is 1.0295. The two ratios are shown in Figure 3 (b) and Figure 4 (b) as black hard line and magenta dashed line. It is clear that most FP improved ratios on unsupervised sub-groups are lower than the magenta dotted line- the minimum TP improved ratio on knowledge-based sub-groups.

6. Conclusions

This paper proposes a JSC-OCC model for real world genuine one class problems. Instead of focusing on the OCC models, we delved into the OCC data-set and developed divided sub-classes and combined sub-classifiers model: the JSC-OCC method. In genuine one class problems, especially in the medical and biological domains, the sub-classes are a natural exploration method. Therefore, the proposed method is very practical in these application areas. The experiments show the results that we expected and indicate improved performances to normally applying the OCC method.

The proposed JSC-OCC method also helps features reduction. Without outlier knowledge, genuine one class problems can only select features by an unsupervised method. With sub-classes identified, the supervised feature selection method can be applied to the sub-class. Most of the unsupervised feature selection or dimension reduction techniques just compress the original feature space into a target feature space and the selected features cannot be explained. This difficulty can also be overcome by applying the JSC-OCC method. Though we change the feature selection object, the experiments show that this feature selected union model will not decrease performance. We conclude that the appropriate sub-class of data-set and features makes the JSC-OCC model perform extremely well.

The proposed JSC-OCC method also has better performance on unsupervised clustering groups if we choose a suitable OCC model, cluster method and cluster number. However, large cluster numbers degenerate the JSC-OCC performances. We also found that the spectral cluster method is better than the K-means cluster method when applying JSC-OCC. The results of the JSC-OCC method on unsupervised sub-clusters cannot compete with our proposed method on knowledge based sub-clusters.

Further research into this method will select appropriate sub-classes. There should be enough cases in the data-set, otherwise there will not be enough cases in the divided the sub-data-set. Sometimes the imbalances in the sub-classes will also affect the CS-OCC method. If there are no sub-classes in the data-set, we will apply the clustering model to cluster the data-set into sub-classes. Though this is not as natural as the more meaningful sub-class, this may be a solution for this kind of problem. The main concern is to find an appropriate cluster number. These issues will be investigated in future research.

Acknowledgments

The work presented in this paper was supported by the Australian Research Council (ARC) under Discovery Project DP088739. Special thanks to Dr Declan Butler for providing an important data-set which was used in this study.

References

1. M. Bramer, S. Crone, J. Guajardoand R. Weber, A study on the ability of Support Vector Regression and Neural Networks to Forecast Basic Time Series Patterns. In *Artificial Intelligence in Theory and Practice*, (Springer Boston, 2006), pp 149-158.
2. S. Vong, B. Coghlan, S. Mardy, D. Holl, H. Seng, S. Ly, M. J. Miller, P. Buchy,Y. Froehlich, J. B. Dufourcq, T. M. Uyeki, W. Limand T. Sok, Low frequency of poultry-to-human H5NI virus transmission, southern Cambodia, 2005. *Emerg Infect Dis* **12** (2006) 1542-7.
3. OIE UPDATE ON HYGHLY PATHOGENIC AVIAN INFLUENZA IN ANIMALS (TYPE H5 and H7). http://www.oie.int/eng/info_ev/en_AI_avianinfluenza.htm (20/11),
4. WHO, Cumulative Number of Confirmed Human Cases of Avian Influenza A/(H5N1) Reported to WHO. 2010.
5. M. Gauthier-Clerc, C. Lebarbenchonand F. Thomas, Recent expansion of highly pathogenic avian influenza H5N1: a critical review. *Ibis* **149** (2007) 202-214.
6. J. D. Brown, D. E. Stallknechtand D. E. Swayne, Experimental infection of swans and geese with highly pathogenic avian influenza virus (H5N1) of Asian lineage. *Emerg Infect Dis* **14** (2008) 136-42.
7. H. Chen, G. J. D. Smith, S. Y. Zhang, K. Qin,J. Wang, K. S. Li, R. G. Webster, J. S. M. Peirisand Y. Guan, Avian flu H5N1 virus outbreak in migratory waterfowl. *Nature* **436** (2005) 191-192.
8. A. M. Kilpatrick, A. A. Chmura, D. W. Gibbons, R. C. Fleischer, P. P. Marraand P. Daszak, Predicting the global spread of H5N1 avian influenza. *Proceedings of the National Academy of Sciences (PANS)* **103** (2006) 19368-19373.
9. T. Weberand N. Stilianakis, Ecologic immunology of avian influenza (H5N1) in migratory birds. *Emerg Infect Dis* **13** (2007) 1139-1143.
10. M. Gilbert, X. Xiao, D. U. Pfeiffer, M. Epprecht, S. Boles, C. Czarnecki, P. Chaitaweesub, W. Kalpravidh, P. Q. Minh, M. J. Otte, V. Martinand J. Slingenbergh, Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. *Proceedings of the National Academy of Sciences (PANS)* **105** (2008) 4769-4774.
11. M. M. Moya, M. W. Kochand L. D. Hostetler One-class classifier networks for target recognition applications, in *World Congress on Neural Networks*, Portland, 1993, pp 797-801.
12. M. Markouand S. Singh, Novelty detection: a review--part 1: statistical approaches. *Signal Processing* **83** (2003) 2481-2497.
13. V. J. Hodgeand J. Austin, A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review* **22** (2004) 85-126.
14. B. Scholkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smolaand R. C. Williamson, Estimating the Support of a High-Dimensional Distribution. *Neural Computation* **13** (2001) 1443-1471.

15. D. M. J. Taxand R. P. W. Duin, Support Vector Data Description. *Mach. Learn.* **54** (2004) 45-66.
16. H. Yu, J. Hanand K. C. Chang, PEBL: Web Page Classification without Negative Examples. *IEEE Transactions on Knowledge and Data Engineering* **16** (2004) 70-81.
17. C. Varun, B. Arindamand K. Vipin, Anomaly detection: A survey. *ACM Comput. Surv.* **41** (2009) 1-58.
18. H. Bu,J. Wangand X. Huang, Fabric defect detection based on multiple fractal features and support vector data description. *Engineering Applications of Artificial Intelligence* **22** (2009) 224-235.
19. C. Sanchez-Hernandez, D. S. Boydand G. M. Foody, One-Class Classification for Mapping a Specific Land-Cover Class: SVDD Classification of Fenland. *IEEE Transactions on Geoscience and Remote Sensing* **45** (2007) 1061-1073.
20. H. Garcesand D. Sbarbaro, Outliers detection in environmental monitoring databases. *Engineering Applications of Artificial Intelligence* **24** (2011) 341-349.
21. T. Onoda, H. Murataand S. Yamada, Non-Relevance Feedback Document Retrieval based on One Class SVM and SVDD. In *International Joint Conference on Neural Networks, 2006 (IJCNN '06)* Vol. (IEEE, Vancouver, Canada, 2006), pp 1212-1219.
22. L. Manevitz, M. and M. Yousef, One-class svms for document classification. *J. Mach. Learn. Res.* **2** (2002) 139-154.
23. X. Zhu,X. Wuand C. Zhang, Vague One-Class Learning for Data Streams. In *Ninth IEEE International Conference on Data Mining, 2009 (ICDM '09)*, Vol. (IEEE, Miami, FL,USA, 2009), pp 657-666.
24. P. Juszczak, D. M. J. Tax, E. P. Kalskaand R. P. W. Duin, Minimum spanning tree based one-class classifier. *Neurocomputing* **72** (2009) 1859-1869.
25. Y. Xuand R. G. Brereton, Diagnostic Pattern Recognition on Gene-Expression Profile Data by Using One-Class Classification. *Journal of Chemical Information and Modeling* **45** (2005) 1392-1401.
26. J. C. Setubal, S. Verjovski-Almeida, E. J. Spinosaand A. C. P. L. F. de Carvalho, Combining One-Class Classifiers for Robust Novelty Detection in Gene Expression Data. In *Advances in Bioinformatics and Computational Biology*, (Springer Berlin / Heidelberg, 2005),pp 54-64.
27. A. Bairoch, S. Cohen-Boulakia, C. Froidevaux, J. Reyesand D. Gilbert, Combining One-Class Classification Models Based on Diverse Biological Data for Prediction of Protein-Protein Interactions. In *Data Integration in the Life Sciences*, (Springer Berlin / Heidelberg, 2008),pp 177-191.
28. M. Yousef, S. Jung, L. C. Showeand M. K. Showe, Learning from positive examples when the negative class is undetermined--microRNA gene identification. *Algorithms Mol Biol* **3** (2008) 2.
29. G. Guptaand J. Ghosh, Robust one-class clustering using hybrid global and local search. In *Proceedings of the 22nd international conference on Machine learning*, Vol. (ACM, Bonn, Germany, 2005).
30. W. Daelemans, B. Goethals, K. Morik, K. Hempstalk, E. Frankand I. Witten, One-Class Classification by Combining Density and Class Probability Estimation. In *Machine Learning and Knowledge Discovery in Databases*, (Springer Berlin / Heidelberg, 2008), pp 505-519.

31. O. Kaynak, E. Alpaydin, E. Oja, L. Xu, D. Taxand K.-R. Müller, Feature Extraction for One-Class Classification. In *Artificial Neural Networks and Neural Information Processing — ICANN/ICONIP 2003*, (Springer Berlin / Heidelberg, 2003), pp 177-177.
32. R. Polikar, Ensemble based systems in decision making. *Circuits and Systems Magazine, IEEE* **6** (2006) 21-45.
33. T. Windeatt, Accuracy/Diversity and Ensemble MLP Classifier Design. *Neural Networks, IEEE Transactions on* **17** (2006) 1194-1211.
34. L. Breiman, Bagging predictors. *Machine Learning* **24** (1996) 123-140.
35. R. E. Schapire, The strength of weak learnability. *Machine Learning* **5** (1990) 197-227.
36. B. Vermaand A. Rahman, Cluster Oriented Ensemble Classifier: Impact of Multi-cluster Characterisation on Ensemble Classifier Learning. *Knowledge and Data Engineering, IEEE Transactions on* **PP** 1-36.
37. F. N. Julia, K. M. Iftekharuddinand A. U. Islam, Dialog Act Classification Using Acoustic And Discourse Information Of Maptask Data. *International journal of computational intelligence and application* **9** (2010) 289-311.
38. L. Rokach, O. Maimonand O. Arad, Improving Supervised Learning by Sample Decomposition. *International Journal of Computational Intelligence and Application* **5** (2005) 37-53.
39. D. M. J. Tax. One-class Classification. Doctoral thesis, (Delft University of Technology, 2001).
40. P. Perner, Concepts for novelty detection and handling based on a case-based reasoning process scheme. *Engineering Applications of Artificial Intelligence* **22** (2009) 86-91.
41. S. Villalbaand, P. Cunningham, An evaluation of dimension reduction techniques for one-class classification. *Artificial Intelligence Review* **27** (2007) 273-294.
42. H. Farvareshand M. M. Sepehri, A data mining framework for detecting subscription fraud in telecommunication. *Engineering Applications of Artificial Intelligence* **24** (2011) 182-194.
43. H. Peng, F. Longand C. Ding, Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27** (2005) 1226-1238.
44. D. M. J. Tax DDtools, the Data Description Toolbox for Matlab. http://homepage.tudelft.nl/n9d04/dd_tools.html