Automated novelty detection in the WISE survey with one-class support vector machines[*]

A. Solarz[1], M. Bilicki[2,1,3], M. Gromadzki[4], A. Pollo[1,5], A. Durkalec[1], and M. Wypych[5]

[1] National Center for Nuclear Research, ul. Hoża 69, 00-681 Warsaw, Poland
  e-mail: `aleksandra.solarz@ncbj.gov.pl`; `bilicki@strw.leidenuniv.nl`
[2] Leiden Observatory, Leiden University, 2333 CA Leiden, The Netherlands
[3] Janusz Gil Institute of Astronomy, University of Zielona Góra, 65-417 Zielona Góra, Poland
[4] Warsaw University Astronomical Observatory, 00-001 Warszawa, Poland
[5] The Astronomical Observatory of the Jagiellonian University, 31-007 Kraków, Poland

ABSTRACT

Wide-angle photometric surveys of previously uncharted sky areas or wavelength regimes will always bring in unexpected sources – novelties or even anomalies – whose existence and properties cannot be easily predicted from earlier observations. Such objects can be efficiently located with novelty detection algorithms. Here we present an application of such a method, called one-class support vector machines (OCSVM), to search for anomalous patterns among sources preselected from the mid-infrared AllWISE catalogue covering the whole sky. To create a model of expected data we train the algorithm on a set of objects with spectroscopic identifications from the SDSS DR13 database, present also in AllWISE. The OCSVM method detects as anomalous those sources whose patterns – WISE photometric measurements in this case – are inconsistent with the model. Among the detected anomalies we find artefacts, such as objects with spurious photometry due to blending, but more importantly also real sources of genuine astrophysical interest. Among the latter, OCSVM has identified a sample of heavily reddened AGN/quasar candidates distributed uniformly over the sky and in a large part absent from other WISE-based AGN catalogues. It also allowed us to find a specific group of sources of mixed types, mostly stars and compact galaxies. By combining the semi-supervised OCSVM algorithm with standard classification methods it will be possible to improve the latter by accounting for sources which are not present in the training sample, but are otherwise well-represented in the target set. Anomaly detection adds flexibility to automated source separation procedures and helps verify the reliability and representativeness of the training samples. It should be thus considered as an essential step in supervised classification schemes to ensure completeness and purity of produced catalogues.

Key words. infrared: galaxies – infrared: stars – galaxies: statistics – stars: statistics – Galaxy: fundamental parameters

1. Introduction

Catalogues of astronomical objects derived from sky surveys serve as a foundation for any subsequent scientific analysis. One of their primary uses is to provide information about statistical properties and spatial distribution of the observed sources, and to identify rare objects, especially those whose presence in the dataset is not expected. Regardless of the aim of the survey, it is important to identify what characteristic properties each class of objects exhibits. This information is crucial in order to separate the desired type of sources from the heap of collected data for further analysis.

The nature of an astronomical object can be determined most reliably by analysing its electromagnetic spectrum. However, even the largest spectroscopic surveys undertaken today, designed to provide detailed information about each observed object, usually cover just a fraction of all the sources available for a given instrument. Photometric observations, on the other hand, are capable of delivering data for many more sources at a significantly faster rate and lower costs.

For photometric data the traditional tool for object separation are colour-colour (CC) and colour-magnitude (CM) diagrams, where various types of objects (like stars and galaxies) appear in separate areas due to differences in observed colours (e.g. Walker et al. 1989; Pollo et al. 2010; Jarrett et al. 2017). Today's largest photometric datasets, such as SuperCOSMOS (Hambly et al. 2001), WISE (Wright et al. 2010), or Pan-STARRS (Chambers et al. 2016), contain of the order of a billion catalogued sources each, which means that even now the traditional ways of dealing with the resulting catalogues by direct human inspection are not practicable. These numbers are expected to grow by orders of magnitude with future experiments such as the LSST[1] or SKA[2], so it is crucial to develop automated methods for source classification in the associated data products. With the advent of self-learning algorithms, the task of source separation can now be dealt with much more efficiently and much more reliably, due to the ability of the algorithm to work in a multidimensional rather than two-dimensional parameter space (e.g. CC or CM diagrams) as is usually the case for human analysis.

Machine learning schemes are now widely used to automatically classify astronomical sources. Owing to automated algorithms, selecting objects of significantly different properties (compactness, colour, etc.) from sky surveys has become quite straightforward (Zhang & Zhao 2004; Solarz et al. 2012; Cavuoti et al. 2014; Shi et al. 2015; Heinis et al. 2016).

---

[1] `www.lsst.org`
[2] `www.skatelescope.org`

However, depending on the nature of the survey, we can expect different types of objects to appear within the field of view. In surveys covering large areas of the sky and reaching deep enough to encompass significant amounts of both Galactic and extragalactic sources, the source separation is usually complicated. In such a case restricting the search to just a few basic classes (e.g. stars, galaxies, and quasars) is not sufficient any more, as the closer we get to the Galactic plane the more diverse objects we can expect to find, planetary nebulae, special types of stars (embedded in envelopes or undergoing catastrophic events), regions of interstellar matter, etc. Moreover, the wavelength regime in which observations are made determines what kind of objects can be expected to appear. For instance, in optical surveys we will find far fewer dust-rich objects than in infrared (IR) ones, while hot stars clearly visible in optical and ultraviolet bands will fade away at longer wavelengths. On the other hand, if any unknown objects are present within the data, their properties should stand out from the crowd of the expected ones. However, detecting these outliers is not straightforward as it is not uncommon for rare objects to mimic the appearance of the well-known ones; for instance, a star and a compact galaxy could both be classified as the same source type based on their angular size only.

In Kurcz et al. (2016) an attempt was made to perform an automated, supervised source classification of IR sources from the all-sky survey conducted by the WISE satellite. The training sample was based on the most secure identifications of SDSS spectroscopic sources divided into three classes of expected (normal) objects: stars, galaxies, and quasars. However, such a standard approach of supervised classification is not designed to correctly handle objects with patterns absent during training, or in other words, anomalous sources. Moreover, especially in low-resolution surveys like WISE, in areas of high observed source density such as low Galactic latitudes, measurements are plagued by effects of overcrowding and therefore blending of objects. In such cases the measured properties of objects can display deviant characteristics and create further training biases. These issues were partly avoided in Kurcz et al. (2016) by removing the data from the lowest Galactic latitudes ($|b| < 10°$ and wider by the Galactic bulge) in order to improve the classification. Nevertheless, both the catalogue of galaxy candidates and especially of the putative quasars obtained there exhibited large-scale on-sky variations resulting from issues with the data themselves, but also – and maybe more importantly – from imperfections of the classification approach. One of the goals of the present study is to examine whether those results could be explained by the existence of unaccounted for anomalous sources in the WISE data, and what the prospects are for improving that classification.

By definition, characteristics of the unexpected sources are not known a priori, rendering the standard multi-class approach inapplicable. Novelty detection schemes offer a solution to these problems, and such methods are designed to recognise cases when a special population of data points differ in some aspects from the data which are used to train the machine learning algorithm. It is common to apply these methods to datasets which contain a large number of examples representing the "normal" populations, but for which data describing the "anomalous" populations are insufficient. In this work objects inconsistent with the training data will be defined as anomalous, as in principle they should display novel/outlying properties in the parameter space. In other words, "unknown" patterns in the target set will manifest themselves in the form of points deviating from the "known" sources. A comprehensive review of this type of methodologies developed for machine learning can be found in Hodge & Austin (2004), Agyemang et al. (2006); and Chandola et al. (2009). According to Hodge & Austin (2004), a user can approach the problem of novelty detection in three different ways. The first is based on unsupervised clustering, where outliers are detected without any previous knowledge about the data. The second approach uses supervised classification, where data have to be prelabelled as normal and unknown. The third method is a mixture of the first two and is referred to as semi-supervised recognition, where the algorithm models the normality of the data, and no knowledge about the true nature of test data is assumed. In this third approach the observer designs an algorithm to create a model of how normal data behave based on a large number of representative examples introduced during training. Next, the algorithm investigates previously unseen patterns by comparing them to the model of normality and searches for a score of novelty. This decision threshold is then used to infer whether the data are behaving in a different manner with respect to the training set or not.

A wide selection of outlier detection algorithms is currently available. Some, based on unsupervised approaches like random forest (Ho 1998), were used by Baron & Poznanski (2017) to find SDSS galaxies with abnormal spectra. Other methods, based on semi-supervised graph-based methods like label propagation and label spreading (described in detail in Chapelle et al. 2006), were used by Škoda et al. (2017, 2016) to identify artefacts and interesting celestial objects in the LAMOST survey.

In the present work we use a knowledge-based novelty detection method designed to create a boundary within the structure of the training dataset, support vector machines (SVM, Vapnik 1995), to show the power of anomaly detection algorithms and discuss how they could improve automatic selection schemes for present and forthcoming surveys covering large areas of the sky. As a case study, we will search for potential new source classes in the WISE dataset.

The paper is organized as follows: in Sect. 2 we present the data and the parameter space used by the SVM algorithm; a description of the one-class SVM algorithm we use can be found in Sect. 3 where we discuss the steps of anomaly detection and describe the training process; Sect. 4 contains the results of the application of those procedures to AllWISE data; a summary and conclusions are given in Sect. 5.

2. Data

In this paper we perform anomaly detection in the Wide-field Infrared Survey Explorer (WISE) data, aiming at improving the early all-sky classification results of Kurcz et al. (2016) and at searching for deviant objects with unexpected properties. Exploration of the publicly available AllWISE catalogue (Cutri et al. 2013), which contains over 747 million sources with photometric information, allows us to test the power of basic artificial intelligence algorithms for anomaly detection in order to obtain information about special objects contained within the dataset. Currently AllWISE is the deepest all-sky dataset available to the public which at the same time provides vast amounts of data that can be used to test automatic schemes of classification and detection of novelty.

The WISE telescope (Wright et al. 2010), launched by NASA in December 2009, has been scanning the whole sky, originally in four passbands ($W1$, $W2$, $W3$, and $W4$) covering near- and mid-IR wavelengths centred at 3.4, 4.6, 12, and 23 μm, respectively. The AllWISE Source Catalogue was produced by combining the WISE single-exposure images from the WISE 4-Band Cryo, 3-Band Cryo, and NEOWISE Post-Cryo survey

phases (Mainzer et al. 2014). The angular resolution of the filters is 6.1″, 6.4″, 6.5″, and 12.0″, respectively, and the sensitivity to point sources at the 5σ detection limit is estimated to be no less than 0.054, 0.071, 0.73, and 5 mJy, which is equivalent to 16.6, 15.6, 11.3, and 8.0 Vega mag, respectively[3].

2.1. Source preselection: WISE × SDSS cross-match

The source preselection and parameter space for the purpose of this study follows that of Kurcz et al. (2016). Namely, we focus on reliable measurements only in the $W1$ and $W2$ channels to maximise the completeness and uniformity of the sample. We thus use AllWISE sources which meet the following criteria: profile-fit measurement signal-to-noise ratios `w1snr` ≥ 5 and `w2snr` ≥ 2; saturated pixel fractions `w1sat` and `w2sat` ≤ 0.1. To ensure that we do not preserve any severe artefacts we also apply `cc_flags`$[1, 2] \neq'$ DPHO′, which excludes sources with diffraction spikes, persistence, halos, or optical ghosts. We emphasise that we do not use the $W3$ and $W4$ channels for the preselection nor for the SVM analysis; the former band will only be employed in the verification phase for CC plots.

In the domain of knowledge-based machine learning it is necessary to create a template for the classification of known objects. Therefore, the training data should representatively sample the underlying distribution of target objects within a given parameter space. In the case of this study, an ideal training set would be constructed from a subsample of securely measured WISE sources with well-defined types. However, since at present such datasets are not available for WISE, to create the basis for the training process an external dataset containing sources of interest is needed. For this purpose we construct the training set by cross-matching the AllWISE dataset with the Sloan Digital Sky Survey (SDSS, York et al. 2000) DR13 (SDSS Collaboration et al. 2016), which provides spectroscopic measurements (Bolton et al. 2012). The SDSS spectroscopic sample includes over 4.4 million sources, where galaxies comprise 59%, quasars 23%, and stars 18%. The cross-matching procedure was performed using a 1″ matching radius and resulted in 3 million common sources, of which galaxies constitute 70%, quasars 12%, and stars 18%. This sample of AllWISE sources with a counterpart in SDSS DR13 spectroscopic will henceforth be referred to as AllWISE × SDSS, and below we provide details on cuts applied to it before the training procedure.

2.2. Parameter space

As in Kurcz et al. (2016) where SDSS DR10 was used, the cross-match between AllWISE and SDSS DR13 practically does not provide galaxies fainter than Vega $W1 = 16$ or $W2 = 16$. For the sake of completeness we thus trim all our catalogues, including the target AllWISE, at these limits; the same applies to any other cuts described below. At the bright end the matched catalogue contains practically only stars, we thus apply additional criteria of $W1 > 9.5$ and $W2 > 9.5$ as otherwise such a population of bright stars would be identified as anomalies by our scheme.

In the earlier related studies by Kurcz et al. (2016) and Krakowski et al. (2016), where a more classical approach to supervised learning was applied, the SDSS-based training sets were purified of sources with problematic redshift measurements according to SDSS parameters such as `zWarning` and `zErr`;

this was done to avoid type misidentifications which could have detrimental effects on multi-class source identification. In the case of outlier detection schemes the aim is to search for sources which do not exhibit patterns learned during training. As such algorithms are not designed to provide distinctions between specific classes but rather to show unexpected sources, the quality of the redshift measurement is of little interest here. Even though the SDSS spectroscopic data themselves may contain anomalous sources, the focus of our study is to find interesting objects within the infrared WISE catalogue. Therefore, even if an objects exhibits deviant properties at optical wavelengths but is otherwise well-detected, it will still be included in our training sample as a known source. For that reason we do not apply any data cleaning on the SDSS spectroscopic database.

As mentioned above, our parameter space was limited to two out of four available WISE passbands: $W1$ and $W2$. This is to ensure as many objects as possible in the final catalogue as the $W3$ and especially the $W4$ filters have much lower sensitivities and a much shorter data acquisition period (limited to the cryogenic phase) which leads to their much lower detection rates than at the shorter wavelengths. The $W3$ and $W4$ passbands are dominated by upper limits and non-detections in the WISE database and using them in our study would lead to losing the majority of objects and introducing severely non-uniform distribution of the sources, and significant biases in the photometry.

To ensure the maximum coverage of the parameter space by known sources, instead of using the $W1$ and $W2$ measurements separately, we employ the $W1$ magnitude and the $W1 - W2$ colour. Even though using flux measurements from each filter separately is mathematically equivalent to employing the colours derived from them (e.g. Wolf et al. 2001), usage of colours can enhance the spread area within the parameter space for the considered objects. Finally, to extend the parameter space, we also use a concentration parameter defined as the difference between flux measurements in two circular apertures in the $W1$ passband in radii equal to 5.5″ and 11.0″ centred on a source:

$$\mathtt{w1mag13} \equiv \mathtt{w1mag\_1} - \mathtt{w1mag\_3}, \qquad (1)$$

used previously by Bilicki et al. (2014, 2016), Kurcz et al. (2016); and Krakowski et al. (2016). It serves as a proxy for morphological information: extended sources will typically have larger `w1mag13` values than point-like ones (see Fig. 1). We emphasise that currently the WISE database does not provide any reliable extended source identifications nor isophotal magnitudes, except for a small subset (~500 000) of objects in common with the 2MASS Extended Source Catalogue and for those in some of the GAMA fields (Cluver et al. 2014; Jarrett et al. 2017).

To summarize this part, the chosen parameter space has three dimensions:

1. $W1$ magnitude measurement;
2. $W1 - W2$ colour;
3. concentration parameter `w1mag13`.

All the WISE magnitudes will be given in the Vega system.

2.3. Quality cuts

To purify the data further, we apply two cuts on the concentration parameter. First, we require that `w1mag13` ≥ 0 to remove objects with measured flux decreasing with increased aperture, which are most likely artefacts of source extraction in high-density areas. This cut removes 3360 sources from the AllWISE × SDSS
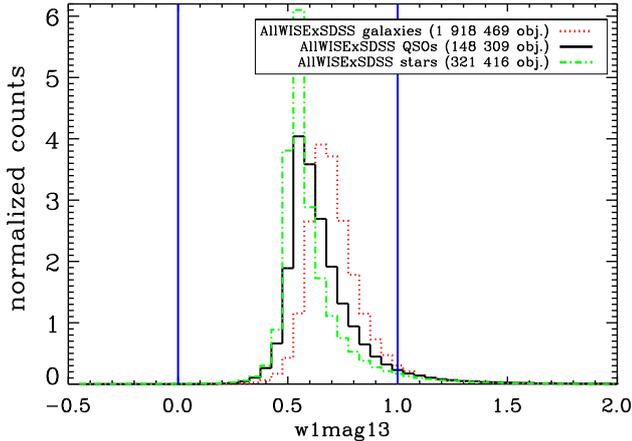
Fig. 1. Distribution of the `w1mag13` parameter (a proxy for flux concentration) for galaxies, stars, and quasars in the cross-match between AllWISE and SDSS spectroscopic sources. Vertical lines mark the position of the applied cuts (`w1mag13 = 0` and `w1mag13 = 1`) which remove blended sources from the training sample.

training set. In our full WISE dataset, this cut eliminates about 600 000 sources, the vast majority of which are located within the Galactic plane and bulge, in Magellanic Clouds, and in M31; these are regions of severe blending in WISE, which is a further confirmation of the spurious nature of these `w1mag13 < 0` objects.

Due to the much lower angular resolution of WISE compared to SDSS (6.1″ in the $W1$ channel vs. 1.3″ in the r band), two objects appearing in close vicinity of one another can be well-separated in SDSS, but may be blended in WISE. This could then lead to high values of the `w1mag13` parameter, suggesting an extended source which in fact is a blend. Such objects would introduce biases during the anomaly detection process. The distribution of the `w1mag13` parameter in our training set is illustrated in Fig. 1. It peaks at ~0.6 for stars and quasars, and at ~0.7 for galaxies. Only a small fraction (2%) of the training sources have `w1mag13 > 1`, and we examined a representative sample of such objects by eye, starting from those with the most extreme values. We have found that the vast majority of them are indeed blends, and this happens even if `w1mag13 ~ 1`. An example is shown in Fig. 2. Two well-separated sources in SDSS (a quasar and a star) are blended in WISE and the 11″ aperture centred on the object of interest gathers a large amount of flux from the second source. Such objects are not usable for the purposes of source separation and anomaly detection despite their usually excellent quality of SDSS measurements. Owing to these considerations, we will not be using objects with `w1mag13 > 1` for training; we then also have to remove such sources from the target catalogue. Such a cut removes a considerable number of AllWISE objects. However, for the purpose of the present analysis, these cuts are necessary as the training sample has to reflect the target sample in terms of parameter ranges. Otherwise target sources with parameter values differing significantly from the input ranges would be automatically marked as anomalous.

After all the cuts discussed above, our training sample includes almost 2.3 million sources, of which 81% are galaxies, 13% are stars and 6% are quasars (see Table 1).

The final AllWISE catalogue that will be used for the novelty detection is composed of 237 million objects; see map in Fig. 3. The most prominent features are our Galaxy and the Magellanic Clouds; however, there is lower surface density in the Bulge, consistent with blending effects in areas of high projected
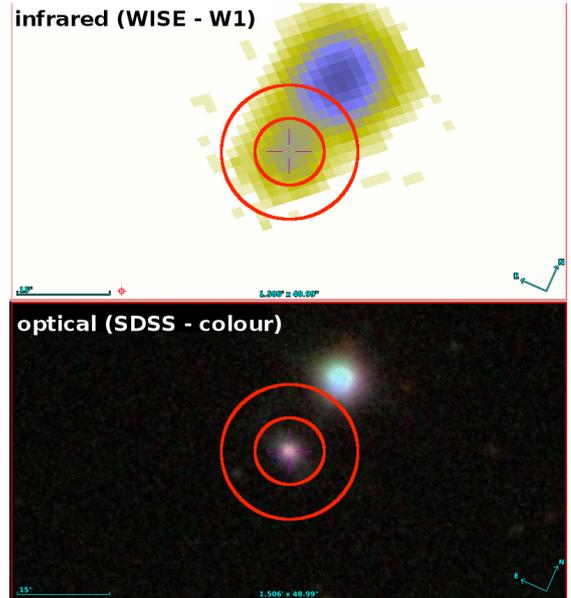
Fig. 2. Example of a quasar with a clean SDSS detection at $\alpha = 192.00$, $\delta = 10.17$ (*lower panel*, colour image constructed form u, g and r fiters) for which the WISE-derived concentration parameter (`w1mag13 ≡ w1mag_1 − w1mag_3`) is 1.13 because of blending with a nearby star (*upper panel*, single-band image with $W1$ flux). Concentric circles mark the aperture in which the `w1mag_1` (5.5″) and `w1mag_3` (11″) magnitudes were measured in WISE.

Table 1. Summary of the training samples of SDSS objects used to train the OCSVM classifier.

SDSS class	N_{obj} before cuts	N_{obj} after cuts
Galaxy	1 918 469	1 827 211
Star	321 416	298 254
QSO	148 309	141 471
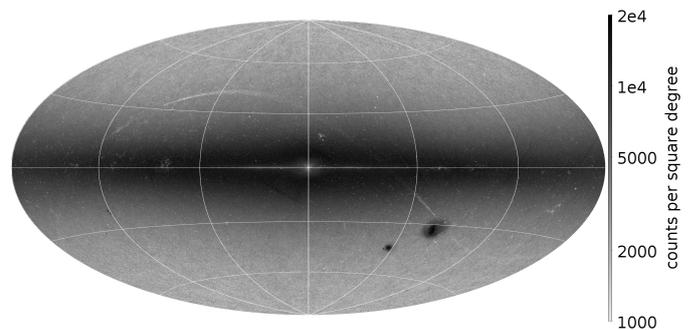
Fig. 3. Sky distribution of the 237 million AllWISE sources used for anomaly detection in this study. See text for details of sample preselection.

density[4]. Also visible are stripes related to WISE instrumental issues[5].

3. Novelty detection

After the introduction of kernel methods (Vapnik 1995; Shawe-Taylor & Cristianini 2004), pattern recognition schemes

---

[4] http://wise2.ipac.caltech.edu/docs/release/allsky/expsup/sec2_2.html
[5] http://wise2.ipac.caltech.edu/docs/release/allwise/expsup/sec2_2.html#w1sat

(ridge regression, e.g. Murphy 2012; Fisher discriminant, e.g. Mika et al. 1999; principle component analysis, e.g. Schölkopf et al. 1999; spectral clustering, e.g. Langone et al. 2015; etc.) have gained in popularity in many branches of science where the amount of data being collected is increasing to the point where human processing is no longer practicable, which is the current situation in astronomy. Kernel methods explore linear and non-linear pair-wise similarity measures. Using non-linear kernels is equivalent to mapping data from the original input space onto a higher dimensional feature space where distinction between patterns can be easier.

This conventional pattern recognition focuses on two or more classes. In a two-class problem we are dealing with a set of training examples $X = (x_i, \omega_i) | x_i \in R^D$, $i = 1...N$, which contain D-dimensional vectors of D characteristic properties (features or observables) for each of the N examples (in astronomy: sources). Depending on the class the object belongs to, it is then given a certain label $\omega = \{-1, 1\}$. Next, out of the training dataset a function $h(x)$ is constructed to estimate which label should be assigned to a new input vector x': $\omega = h(x'|X)$:

$$h(x'|X) : R^D \to [-1, 1]. \tag{2}$$

In the case of classification schemes of more than two classes it is typical to decompose the problem into multiple binary problems. The final classification result combines partial outcomes of binary classifiers by a ranking method. This conventional way, however, ignores any new/outlying data that do not belong to the considered classes. Without any freedom, the algorithm is forced to classify a source as one of the predefined classes, even if it does not fit to any presented category, for example objects that do not occur in an optical-based training sample but are detected in the IR.

To tackle the problem of novel data detection it is possible to modify the standard supervised classification scheme to *one-class classification*. Here, the main class composed of normal/expected data points will be detected separately from all the other data points. In the usual approach to novelty detection it is assumed that the normal class is well sampled, while the outlying class is undersampled. A model of normality $N(\theta)$ (not to be confused with normal distribution), where θ is a free parameter of the model, is deduced and used to assign the novelty scores $n(x)$ to the previously unseen data x. In this sense, increasing scores can be understood as increasing deviation of the points from the normality model. We define the normality threshold as $z(x) = k$ in a way that an example x will be classified as normal if $z(x) \leq k$ or as deviant in the opposite case. Therefore $z(x) = k$ defines the decision boundary. In this way the possibility of misclassification of the objects missing in the training sample is very low, as they will occur simply as deviations from normal.

To search for anomalies within the AllWISE data we have chosen to use a semi-supervised method belonging to knowledge-based algorithms (e.g. Schölkopf et al. 2000) – the support vector machines – as these approaches focus on creating the decision boundary to contain the normal datapoints and are sensitive to outliers in both the training and test set. On the other hand, they do not depend on the distribution of the data within the training set; however, knowledge-based approaches to novelty detection have one drawback: complexity associated with the computational time of kernel functions (see Sect. 3.1). Nevertheless, present-day technology coupled with parallelised computational capabilities significantly shortens the proper kernel choice for a given dataset and corresponding calculations.

There are also several other algorithms for novelty detection, such as reconstruction-based techniques such as neural networks (e.g. Hawkins et al. 2002; Markou & Singh 2003) or subspace-based methods (e.g. Jolliffe 2002; Hoffmann 2007) that model the underlying data and reconstruct an error defined as a distance between the test vector and output of the system, which is then translated to a novelty score. However, even though reconstruction methods offer a flexible way to deal with high dimensionality of the data, they require a predefinition of parameters to define the structure of the model, which leads to two basic problems; the first is the selection of the most effective training method to enable the integration of new units into the existing structure and the second is the need to add a priori information about the saturation point (when no more new units can be added).

Another large family of novelty detection techniques are distance-based approaches, which do not require any a priori knowledge about the data distribution, like nearest neighbour-based techniques (e.g. Hautamaki et al. 2004; Angiulli et al. 2009) or clustering-based techniques (e.g. Yang & Wang 2003; Basu et al. 2004). However, they require a definition of distance metrics to establish similarity between data points, which becomes an increasingly persisting problem especially when dealing with high dimensionality of the parameter space (e.g. Kriegel et al. 2009) as distance measures in many dimensions lose ability to differentiate between normal and outlying data points. Moreover, these methods lack the flexibility of parameter tuning, making the methods unsuitable for full automation.

Owing to the above reasons, we chose to use support vector machines for our study; a detailed description of our approach follows.

3.1. Support vector machines

Support vector machines is one of most commonly used conventional classification algorithms in astronomy. The idea of SVM is based on structural risk minimization (Vapnik & Chervonenkis 1974). For many applications SVM have shown better performance and accuracy than other learning machines and have been used in many branches of astrophysics to solve classification problems and build catalogues (e.g. Beaumont et al. 2011; Fadely et al. 2012; Małek et al. 2013; Solarz et al. 2015; Kovács & Szapudi 2015; Heinis et al. 2016; Marton et al. 2016; Kurcz et al. 2016; Krakowski et al. 2016). Support vector machines maps input points onto a high-dimensional feature space and finds a hyperplane separating two or more classes with as large a margin as possible between points belonging to each category in this space. Then the solution of the best hyperplane is composed of input points laying on the boundary called support vectors (SVs).

Here we outline the basis of the SVM theory in application to classification schemes. Training of an SVM algorithm starts with having a set of observations with labels $(y_1, x_1), ..., (y_l, x_l)$, where $x_i \in R^N$ belongs to one of two classes and has a label $y_i \in \{-1, 1\}$ for $i = 1, ...l$. Each point should contain a vector of features, characteristic values which describe it. Then the algorithm maps each vector from the input space X onto a feature space H using a non-linear function $\Phi : X \to H$. The desired separation plane $w \cdot z + b = 0$ is defined by the pair (w, b) in such a way that each point x_i is separated according to a decision function

$$f(x_i) = sgn(w \cdot z_i + b), \tag{3}$$

where $w \in H$ and $b \in R$. In principle, it is not the explicit knowledge of the mapping function Φ that is needed, but the dot product of the transformed points $\langle \Phi(x_i), \Phi(x_j) \rangle$ (Cortes & Vapnik 1995). Therefore, instead of working with Φ it is possible to
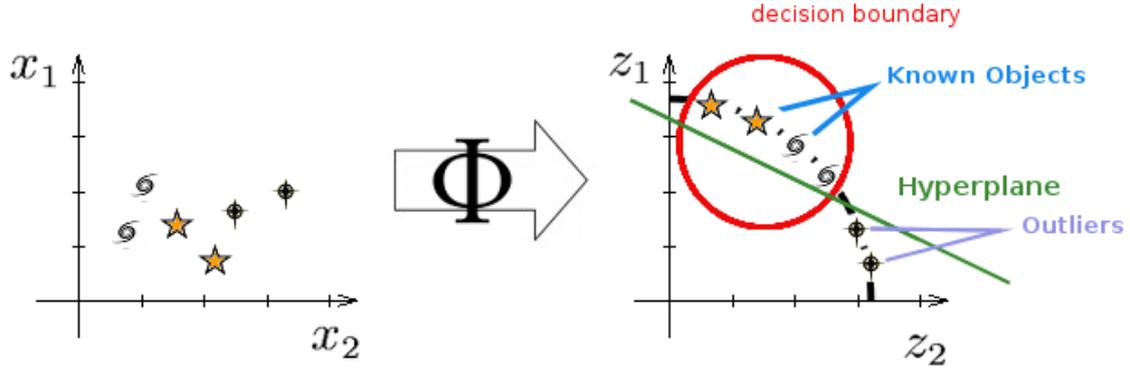
Fig. 4. Schematic representation of OCSVM using an example of the default radial basis kernel. The presented case of classification shows the tightest decision boundary which envelopes the known data (red circle) which can be treated as finding a separating hyperplane in the traditional SVM sense (green line). Unknown objects fall outside the sphere and are marked as outliers.

work with $K:X \times X \rightarrow \boldsymbol{R}$, where K takes two points as input and returns a real value representing $\langle \Phi(x_i), \Phi(x_j) \rangle$. The only condition is that Φ exists if and only if K (called *kernel*) is positive definite (satisfies Mercer's condition; Mercer 1909). Therefore, any function which meets this criterion can be a kernel function. The most commonly featured kernel functions are linear, sigmoid, radial basis, and polynomial, which we describe in more detail in Sect. 3.3.

3.2. One-class SVM reformulation

Schölkopf et al. (2000) introduced an extension of the SVM methodology to pattern recognition as an open set problem. Unlike the traditional SVM algorithm, which is designed to differentiate between classes contained within a given set, *one-class SVM* (hereafter OCSVM) recognizes patterns in a much larger space of classes, unseen in training but which occur in testing. For that purpose, in the absence of a second class in the training data, the algorithm defines an "origin" by mapping feature vectors onto a feature space through an appropriate kernel function and then separates them by a hyperplane with a maximum margin with respect to the origin. The resulting discriminant function is trained to assign positive values in the region surrounding the majority of the training points and negative elsewhere. Hyperplane parameters are derived by solving a quadratic programming problem

$$\text{minimize} \left(\frac{1}{2} \boldsymbol{w} \cdot \boldsymbol{w} + \frac{1}{\nu l} \sum_{i=1}^{l} \xi_i - \rho \right) \qquad (4)$$

subject to

$$(w \cdot \Phi(x_i)) \geq \rho - \xi_i; \ i = 1, 2, ..., l; \ \xi_i \geq 0, \qquad (5)$$

where w and ρ are parameters of the separation hyperplane, Φ is the mapping function of the input parameter space to a feature space, ν is the asymptotic fraction of outliers (anomalies) allowed, l is the number of training points, and ξ is a slack variable which penalizes misclassifications. The decision function $f(x) = sgn(w \cdot \Phi(x) - \rho)$ determines point labels (e.g. +1 for known instances and −1 for novel points). A schematic idea behind OCSVM is shown in Fig. 4.

In this approach the parameter ν is interpreted as the asymptotic fraction of data labelled as outliers. The choice of the outlier fraction ν implies that the knowledge about the frequency of appearance of novel points is known a priori (e.g. Manevitz & Yousef 2007). Otherwise, the value has to be tuned

as a free parameter together with other unknowns. It is worth noting that domain-based approaches, such as this one, regulate the position of the novelty boundary using only those data with the closest proximity to it and that the properties of the distribution of data in the training set have no influence on this process (e.g. Tax & Duin 1999; Le et al. 2010, 2011; Liu et al. 2011). The only drawback of the presented method is the complexity associated with the choice and computation of the kernel functions. Moreover, the parameters controlling the size of the boundary area should be properly adjusted, increasing the computational time (e.g. Tax & Duin 2004). In this work we use the R (R Core Team 2013)[6] implementation of SVM included in the `e1071` package (Meyer et al. 2015), which provides an interface to `libsvm`. We use `doParallel`[7] and `caret`[8] packages to parallelize the computations

3.3. Classification scheme

To make a selection of outlying data it is crucial to create the best-suited classifier for a given dataset. In our application, the classifier is trained on sources with spectroscopic measurements in the SDSS database treated as a single class, and present also in the AllWISE catalogue. For this purpose we include all sources from the AllWISE × SDSS cross-match in the training sample. Unlike in the case of classical SVM, the imbalance of the training set has no influence on the OCSVM training, as we create only one known class. The quantity and ratios of specific classes are not an issue here. For that reason OCSVM can also be treated as an alternative approach to dealing with imbalanced datasets which offers no information loss during training (e.g. Batuwita & Palade 2013). With the training sample selected, the algorithm has to be trained to recognise the normal patterns, which in the case of OCSVM means finding the best-suited volume encompassing the training points, which will later be used as a decision boundary between what the algorithm finds as normal and deviating patterns. This procedure involves searching for the most appropriate kernel function, which governs the topology of the surface enclosing the training sample. To ensure the best performance of the algorithm it is necessary to choose an appropriate kernel function for the given training set and to find its meta-parameters to train the novelty detector which

---

[6] http://www.R-project.org
[7] https://cran.r-project.org/web/packages/doParallel/index.html
[8] https://cran.r-project.org/web/packages/caret/index.html

will best suit the input data. As no two datasets are the same, it is natural that there is no universal kernel function optimal for each classification problem. This makes testing several functions a vital step in any kernel-based machine learning process (Sangeetha & Kalpana 2010).

In this study we test four basic shapes of kernel functions in the application of the novelty detection to the AllWISE dataset

i) linear kernel: $u^T v$,
ii) sigmoid kernel: $\tanh(\gamma u^T v + C)$,
iii) radial basis kernel: $\exp\left(-\gamma \|u - v\|^2\right)$,
iv) polynomial kernel: $(\gamma u^T v + C)^d$,

where u and v are vectors in the input space, $\|\cdot\|^2$ is the squared Euclidean distance between the two feature vectors, γ is a scaling parameter, d is the degree of the polynomial function, and C is a constant. These meta-parameters need to be tuned for each given dataset; an exception is the linear kernel which does not have any free parameters.

Taking into account all the above points, the following steps need to be applied for each considered kernel function in order to determine the best-fitting topology of the separation hypersurface for any dataset:

1. *Division of the training set*:
    The full training set is divided into two subsets, where one is used for the actual training and the other is used as a validation subset to verify the accuracy of the created hypersurface against sources with known class which were not used by the algorithm to find the model. We create the training set out of a random 99% of known objects; the validation set contains the remaining 1% of known sources. The percentage left for validation is small but sufficient to verify whether the classifier works well on previously unseen data.
2. *Wide grid search*:
    Training a classifier means finding the best set of kernel function meta-parameters. They are determined by searching through a loosely spaced grid of meta-parameter values describing each kernel (e.g. d, γ, and C in the case of the polynomial kernel) and the ν parameter specifying the expected outlier fraction.
3. *Estimation of training accuracy:*
    For each tested combination of meta-parameters we count how many times an object with known nature was correctly classified by the OCSVM (true positive; TP) and how many times a known object was classified as an outlier (false negative; FN). Based on these counts accuracy is calculated as $acc_{\text{train}} = TP/(TP + FN)$. Moreover, we count the number of SVs used to find the decision boundary. The fewer the points treated as SVs, the better: when the data is well-structured only a small number of SVs are used; all the remaining training points will not be used in the calculation of the final boundary. High numbers of SVs mean that the topology of the surface is complex, and that the data cannot be easily contained within the boundary.
4. *Estimation of validation accuracy:*
    The algorithm, trained in the previous step, with its best-suited meta-parameters of its kernel, is applied to classify the validation set. Thanks to the knowledge of the true labels of this set it is possible to verify how well the trained algorithm is working on previously unseen data by calculating $acc_{\text{valid}} = TP/(TP + FN)$ and therefore to estimate how well it will work on truly unknown data. As above, the number of SVs is also taken into account here.

Table 2. Input parameter ranges for the grid search for the kernels tested in Sect. 4 (see text for details).

Kernel	ν	γ	C	d
Linear	0.0001–0.69	–	–	–
Radial	0.0001–0.69	0.001–10 000	–	–
Sigmoidal	0.0001–0.69	0.001–10 000	0–4	–
Polynomial	0.0001–0.69	0.001–10 000	0-4	2–3

5. *Fine-tuning of grid search*:
    We tighten the grid search around the best values from the wide grid to fine-tune the meta-parameter choice for the best performance (by repetitive measurements of acc_{train}, acc_{valid}).

The search for the free parameters of each kernel is done within reasonable expected ranges. To find these ranges we follow the scheme of Chapelle & Zien (2005), where it is first necessary to fix initial values for each set of parameters which will provide the most reliable orders of magnitude. In the case of the one-class problem we use the median of pairwise distances of all training points as the default for γ. The default for ν is taken as the inverse of the empirical variance s^2 in the feature space calculated as $s^2 = \frac{1}{n}\Sigma_i K_{ii} - \frac{1}{n^2}\Sigma_{i,j} K_{i,j}$ from an $n \times n$ kernel matrix K. For the degree of the polynomial we consider only two possible values, 2 and 3, as higher degrees would create a boundary whose topology is too complex, which would result in overfitting the model. Then we use multiples (10^k for $k \in \{-3, ..., 3\}$) of the default values as the grid search range.

After pinpointing the best set of parameters for each kernel function we perform fine-tuning of the grid search, where we search the grid around those best parameters with a much smaller step (multiples of 2^k). However, we find that the maxima of the performance around the most optimal parameters found on the sparser grid are very broad and the fine-tuning of the grid does not improve the performance of the classifier. Ranges of searched parameter values for each kernel are summarized in Table 2.

Upon finalization of the training and verifying that the classifier performs on a satisfactory level, it is applied to the AllWISE target data to search for true outlying objects.

4. Results of OCSVM application to AllWISE data

In this section we present the results of applying the OCSVM algorithm to the AllWISE catalogue. Following the discussion presented in the previous section, we started by determining the most appropriate kernel for this training set. We found that in this case the preferred kernel function is radial-based with optimal parameters $\gamma = 0.1$ and $\nu = 0.001$ as it provides the highest training and validation accuracies and is characterized by the smallest number of SVs. The optimal parameters and training/validation performance of all four tested kernels are summarised in Table 3.

Having determined the proper kernel, we trained the OCSVM anomaly detector on the AllWISE × SDSS training set, and subsequently applied it to AllWISE data preselected as described in Sect. 2. As a result, we obtained a sample of 642 353 sources classified as unknown by the algorithm. We show their sky distribution in Fig. 5; as is obvious from the plot, the vast majority of the sources is located within the Galactic Plane and Bulge (90% are within $|b_{\text{Gal}}| < 15°$) and in other confusion areas: Magellanic Clouds, Galactic dust clouds, and even M31 and M33. This is to be expected as the 6″ spatial resolution of the WISE satellite leads to severe blending in areas of

Table 3. Performance of the kernel functions tested in the training of the OCSVM algorithm.

Kernel	Best parameters				N_{SV}	acc_{train}	acc_{valid}
	ν	γ	C	d			
Linear	0.1	–	–	–	892	27.02%	31.33%
Sigmoid	0.001	1	0	–	125	99.99%	98.80%
Radial	0.001	0.1	–	–	48	99.99%	99.98%
Poly	0.005	100	0	3	53	97.87%	96.66%
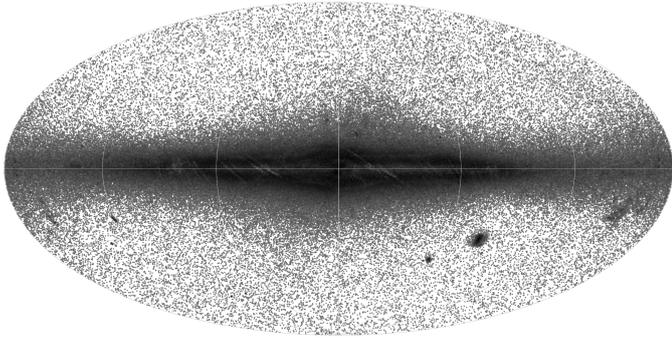
Fig. 5. Sky positions of all the objects classified as unknown in the application of the OCSVM anomaly detector to the AllWISE catalogue (642 353 sources), shown in Galactic coordinates.

high projected density, which in turn results in anomalous (spurious) photometric properties of these blended objects. However, as discussed below, except for such artefacts, our anomaly detector also flagged a considerable number of genuine sources of astrophysical interest.

To gain insight into the nature of these anomalies, we started by looking at their WISE colours. It is important to note that the OCSVM algorithm itself does not provide any means of discriminating various populations among the outliers. Relying on the colours to identify the groups in the resultant anomalies is the most basic approach. It is possible to refine this task by employing clustering algorithms (e.g. Han et al. 2011) which could find different classes within the outlier group. For the bright sources the problem can also be approached by using passband images directly (e.g. Hoyle 2016), but the speed of data processing would significantly decrease. However, in this work we restrict ourselves to a first look at the anomalies, and we mostly use a single WISE colour, $W1 - W2$, for that. In Fig. 6 we present their $W1 - W2$ distribution and compare it to the training set, divided according to SDSS source classes (stars, galaxies, quasars). We observe multi-modal behaviour of this colour for the detected anomalies, with three peaks at $W1 - W2 \sim -1$, ~ -0.5, and ~ 1.7. The peaks are separated by minima at $W1 - W2 = -0.65$ and at ~ 0.8. It is interesting to note that the latter is the same as the WISE active galactic nuclei (AGN) separation criterion first proposed by Stern et al. (2012); we discuss these red sources in more detail in Sect. 4.2. The total number of sources contained in each considered group is shown in Table 4.

This division into roughly three groups is also confirmed in the CC diagram where the $W2 - W3$ WISE colour is used as the second dimension (Fig. 7). To construct this diagram we used only those sources which had positive signal-to-noise ratio in the $W3$ band, which is 38% of the full anomaly sample. In addition, for objects with $0 < \text{w3snr} < 2$, which have only $W3$ upper limits in the WISE database, we applied the correction of +0.75 mag as discussed in the Appendix of Krakowski et al. (2016). This CC diagram gives indications of
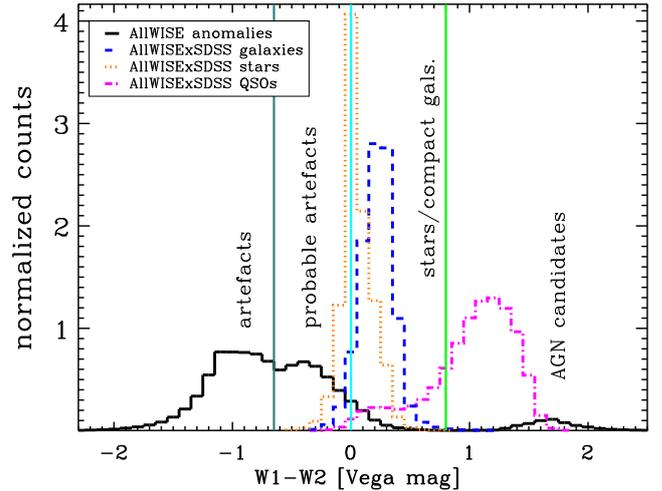
Fig. 6. Distributions of the $W1 - W2$ colour for AllWISE anomalies found in this work (solid black) compared with known sources from AllWISE × SDSS used to train the OCSVM algorithm. Galaxies (1 827 241 objects) are marked by blue dashed lines; stars (298 269 objects) by orange dotted lines; and quasars (141 494 objects) by magenta dot-dashed lines. Vertical lines mark the colour cuts applied to the OCSVM anomaly sample dividing it into four possible subgroups.

Table 4. Number of sources identified as anomalies by the OCSVM algorithm, further divided according to the $W1 - W2$ colour cuts.

Colour cut	N_{obj}
$W1 - W2 < 0$	575 598
$0 \leq W1 - W2 < 0.8$	26 990
$W1 - W2 \geq 0.8$	39 940

the nature of the detected anomalies: according to Fig. 26 in Jarrett et al. (2011) or Fig. 5 in Cluver et al. (2014), stars concentrate around $W1 - W2 \simeq W2 - W3 \simeq 0$, elliptical galaxies have $W1 - W2 \gtrsim 0$, and $0.5 < W2 - W3 < 1.5$ while spirals span $0 < W1 - W2 < 0.5$ and $1 < W2 - W3 < 4.5$; quasars are much redder in $W1 - W2$ than most of inactive galaxies while their $W2 - W3$ is similar to that of some spirals. There are also some more specific sources located on that diagram, such as (U)LIRGs or brown dwarfs, but here we will restrict ourselves to the basic three classes (stars, galaxies, and quasars) in our basic division of the anomalies. Compared with the theoretical $W1 - W2$ vs. $W2 - W3$ diagram, we see that the upper cloud of our anomalies is located roughly at the (obscured) AGN locus, while the two lower ones do not seem to be consistent with any normal sources in this plane.

Below we discuss in more detail the possible nature of these three groups of sources. We reiterate that as most of the detected anomalies do not have measurements in the $W3$ band, the distinction will be made only based on their $W1 - W2$ colour.

4.1. Anomalies with extremely low $W1 - W2$ colour: photometric artefacts

We begin our detailed investigation into the nature of the detected anomalies by looking at those with extremely low $W1 - W2 < -0.65$. This is the majority (55%) of the outliers found by OCSVM. Their sky distribution (Fig. 8), and the fact that they have such an extremely blue and unphysical $W1 - W2$ colour, is a clear indication that these are sources with spurious photometry due to blending. In what follows we will not deal with these
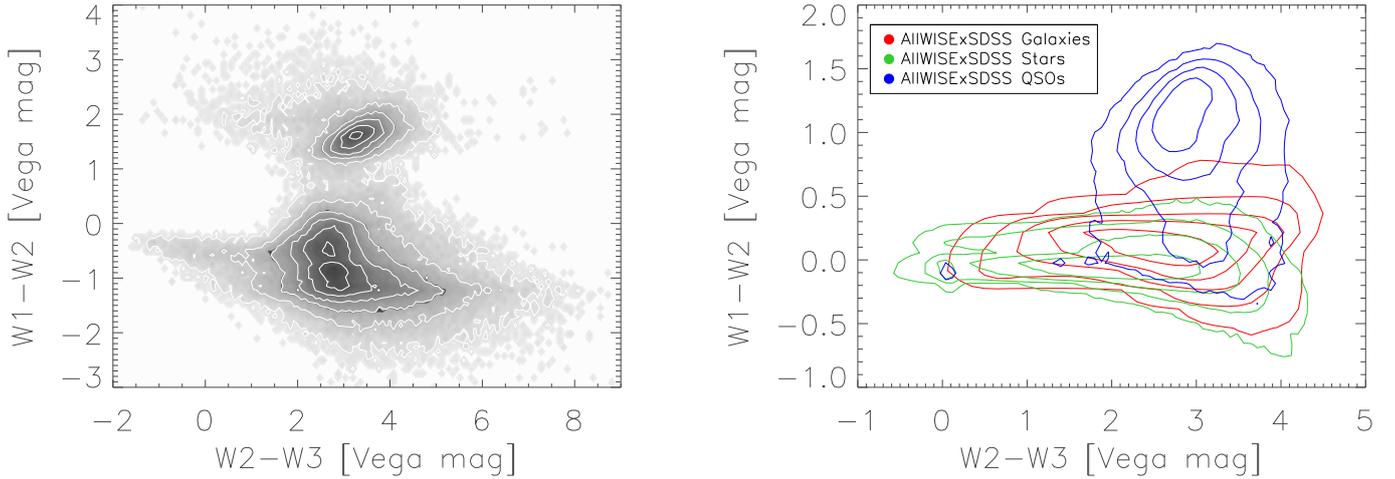
Fig. 7. *Left panel*: WISE colour–colour ($W1 - W2$ vs. $W2 - W3$) diagram for the sources identified by OCSVM as outliers in AllWISE. The plot shows only sources detected in the $W3$ band (188 496 objects comprising 38% of all the anomalies). The grey scale marks the density of displayed points in linear bins. *Right panel*: $W1 - W2$ vs. $W2 - W3$ diagram for AllWISE × SDSS sources used for OCSVM training. We note the different ranges of the axes in the two panels.
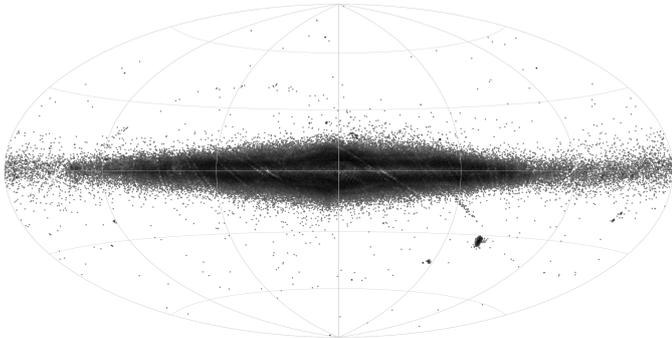
Fig. 8. Sky distribution of AllWISE anomalous sources with $W1 - W2 < -0.65$ (354 301 objects) shown in Galactic coordinates.

Fig. 9. Difference between *Spitzer* IRAC $I2$ and WISE $W2$ magnitudes for the anomalous sources with $W1 - W2 < -0.65$ (354 301 sources) as a function of the WISE $W1 - W2$ colour. The $I2$ measurements were taken from the GLIMPSE survey of the Galactic Plane. Contours mark the density of the displayed points in linear bins.

objects further. We note, however, that their identification by our anomaly detector is evidence of its successful operation.

Even after removing the $W1 - W2 < -0.65$ artefacts, there is still a significant fraction of sources with a very negative and most likely non-physical $W1 - W2$ colour (cf. Fig. 6). Almost all such sources are confined to the Galactic plane and bulge; to inspect whether they have indeed spurious photometry, we checked their properties in GLIMPSE (Benjamin et al. 2003)[9], a *Spitzer* survey covering the inner Galactic Plane and Bulge within $|b_{Gal}| < 5°$ & $|l_{Gal}| < 65°$. As filter coverage of *Spitzer* IRAC $I1$ (centred at 3.6 μm) and IRAC $I2$ (centred at 4.5 μm) are very comparable to WISE $W1$ and $W2$, respectively (Jarrett et al. 2011), this is an adequate test to confront corresponding flux measurements from the GLIMPSE catalogue with those of the OCSVM AllWISE anomalies.

Our sample of outliers with $0 > W1 - W2 > -0.65$ has almost 17 000 counterparts in GLIMPSE within a 3″ matching radius. By comparing the IRAC $I1$ & $I2$ vs. WISE $W1$ & $W2$ measurements, we found that for the shorter-wavelength channels the IRAC and WISE magnitudes match very well, while in the case of $I2$ vs. $W2$ comparison there is a clear discrepancy: WISE measurements in this band significantly underestimate the fluxes with respect to IRAC. What is more, this bias increases with decreasing $W1 - W2$ colour (Fig. 9). Sources

with $0 > W1 - W2 > -0.65$, however, could hide a fraction of real sources of astronomical interest as objects with moderately negative $W1 - W2$ colours have been reported (e.g. Banerji et al. 2013; Jarrett et al. 2017). Current parameter space does not allow for a proper distinction between the real and spurious sources within this anomaly group; only a more in-depth analysis with clustering algorithms could reveal more insight on that matter. As this task is beyond the scope of this work, in the present approach we restrict ourselves to treat all anomalies with $W1 - W2 < 0$ either as having spurious $W2$ photometry or as problematic, and we also remove them from further examinations. This cut only affects confusion areas (Galaxy, Magellanic Clouds) and significantly purifies the anomaly dataset.

4.2. Anomalies with $W1 - W2 > 0.8$: AGN/quasar candidates

We now turn our attention to the group of sources with $W1 - W2 > 0.8$, which clearly stand out in the CC diagram (Fig. 7); there are almost 40 000 such anomalies in our sample. As already

---

[9] We used GLIMPSEII 2.1 Data Release, http://www.astro.wisc.edu/glimpse/glimpse2_dataprod_v2.1.pdf
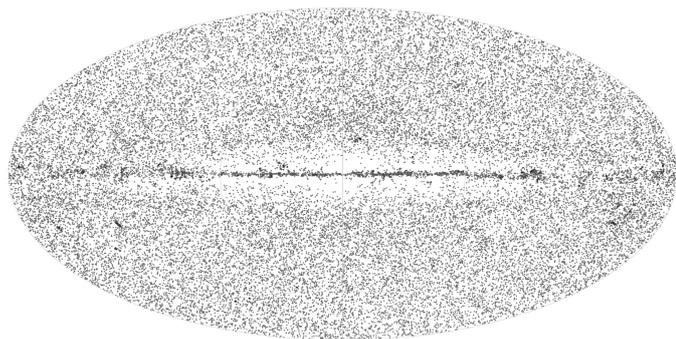
Fig. 10. Sky distribution of AllWISE anomalous sources with $W1 - W2 \geq 0.8$ (39 940 objects).

Fig. 11. $W1$ magnitude distributions for the three main types of anomalies identified by OCSVM in AllWISE: with $W1 - W2 < 0$, solid black (575 423 sources); with $0 \leq W1 - W2 < 0.8$, orange dotted (26 990 sources); with $W1 - W2 \geq 0.8$, blue dashed (39 940 sources), compared with the known sources from AllWISE × SDSS used to train the OCSVM algorithm: galaxies, magenta dot-dashed (1 827 241 sources); stars, red dashed (298 269 sources); and quasars, cyan triple-dot-dashed (141 494 sources).

mentioned, their location in this diagram is consistent with them being AGN/QSO, i.e. high-redshift extragalactic sources. There are several lines of evidence supporting this hypothesis.

First of all, these sources are very uniformly distributed over the entire sky (Fig. 10), and are preferentially located outside the Galactic Bulge, except for some at the Galactic equator (5200 within $|b_{\mathrm{Gal}}| < 3°$) which must be artefacts of WISE blending in a similar way as the very low $W1 - W2$ sources of Sect. 4.1. Secondly, these anomalies are mostly faint: their $W1$ counts peak at the limit of the catalogue used here, $W1 = 16$ (see Fig. 11). Furthermore, almost all of them (over 95%) have WISE detections in the $W3$ channel (w3snr > 0). We reiterate that this channel was not used in the source preselection procedure; thus, as its sensitivity in WISE is much lower than that of the two shorter-wavelength bandpasses, this indicates that these sources are intrinsically bright at (observed) 12 μm. As we can quite safely exclude the situation in which stellar light would be redshifted to this channel (one would need $z > 2$, which for $W1 < 16$ would mean intrinsic brightness of ~ -30 mag or brighter at rest-frame $\lambda \sim 1$ μm), this points to emission from dust as 12 μm observations are sensitive to warm dust radiation (e.g. Sauvage et al. 2005) and polycyclic aromatic hydrocarbon emission lines (PAH, e.g. Brandl et al. 2006) at redshifts lower than 2. On the other hand, a cross-match with the all-sky 2MASS data (Skrutskie et al. 2006) gave only 1500 sources, most of which in its *Point* Source Catalogue (PSC) and just a handful (45) are extended (in 2MASS XSC, Jarrett et al. 2000). This shows that most of our QSO candidates are not in the local volume, as 2MASS provides a very complete census of the local Universe (Bilicki et al. 2014; Rahman et al. 2016).

Further insight into the nature of these objects is gained by checking for their presence and properties in the SDSS photometric catalogue[10]. For training we used only sources with SDSS spectroscopy, while the general photometric dataset from Sloan is obviously much larger and more complete at a price of much more limited information on the real nature of the detected sources. By cross-matching with SDSS photometric, we found about 7000 of our AGN candidates to have counterparts there within a matching radius of 1″. About 3000 of them are also present in the SDSS DR12 photometric redshift catalogue of Beck et al. (2016), which contains SDSS-resolved galaxies up to $z = 1$ – and these matched objects have mean $\langle z \rangle \sim 0.5$. By extrapolating to the full extragalactic sky, this exercise suggests that about 40% of these anomalies would have no optical counterpart in an SDSS-depth all-sky catalogue if one existed. On the other hand, about 25% of these AGN candidates seem to be residing in optically resolved galaxies at $z < 1$.

According to studies of AGNs identified in both WISE and SDSS (e.g. Yan et al. 2013; Donoso et al. 2014), the combined optical-MIR $r - W2$ colour can be used as a diagnostic to differentiate between unobscured/type-1 and dust-obscured/type-2 AGN/QSO candidates, the division being $r - W2 \sim 6$ (both Vega). We have thus checked the behaviour of this colour in our sample of anomalies with $W1 - W2 > 0.8$ also present in SDSS. Indeed, we observe bimodality in the $r - W2$ colour with the division roughly at $r - W2 \sim 6$ (Fig. 12); a similar bimodality is also present in the distribution of the $r - W1$ colour. This suggests that the OCSVM selects both types of AGN populations, although we emphasise that their $W1 - W2$ colour is in most cases much redder than for the AllWISE × SDSS spectroscopic QSOs, used as part of the training (Fig. 13).

In the next step we compared our findings with other WISE-based QSO candidate selections. We will limit ourselves to those works which presented all-sky WISE data in this context, namely Secrest et al. (2015) and Kurcz et al. (2016). The OCSVM-selected QSO candidates have much redder IR colours than the QSO candidates of Kurcz et al. (2016), but also redder than the quasars in the SDSS-based training sample used both here and in that paper (cf. Fig. 13). In addition, they do not show the anomalous sky distribution present in Kurcz et al. (2016, non-uniform distribution on the sky with somewhat larger surface density close to the ecliptic than at the ecliptic poles). There are over 22 000 common sources between the OCSVM-QSO sample and the Kurcz et al. (2016) AllWISE QSO candidate one, and the common sources have $W1 - W2 < 1.8$. Beyond those values, however, OCSVM selects much redder QSO candidates, missed by the classical approach of source classification: in Kurcz et al. (2016) the very high $W1 - W2$ objects were assigned to random classes and had roughly equal probabilities of belonging to any of them.

Finally, we compared the results of the OCSVM QSO selection with the publicly available AllWISEAGN catalogue (Secrest et al. 2015) containing over 1.4 million AGN candidates extracted from AllWISE following the formulae of Mateos et al. (2012). In addition to another method of AGN/QSO selection

---

[10] Avaliable at http://skyserver.sdss.org/dr13
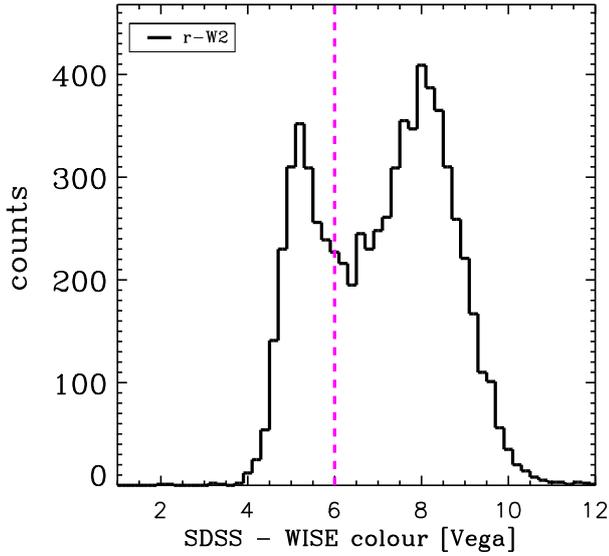
Fig. 12. Histogram of the $r - W2$ colour for the AGN candidates identified by OCSVM which have a counterpart in the SDSS photometric catalogue (7053 sources). The vertical line marks the division between obscured and unobscured AGNs according to Donoso et al. (2014). The SDSS r-band magnitudes were converted to the Vega system following Blanton & Roweis (2007).
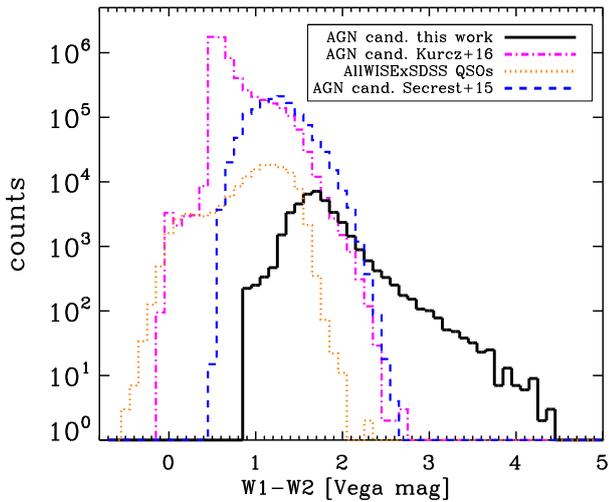
Fig. 13. Histograms of the $W1 - W2$ colour for quasars and quasar candidates from the following datasets: AllWISE paired up with SDSS DR13 spectroscopic (dotted orange; 141 494 sources; SDSS Collaboration et al. 2016); AllWISE OCSVM AGN candidates (black solid line; 39 940 sources; this paper); AllWISE SVM AGN selection (magenta dot-dashed; 4 443 962 sources; Kurcz et al. 2016); and AllWISEAGN (blue dashed; 1 354 775 sources; Secrest et al. 2015).

– colour-based vs. automated – the AllWISEAGN sample also uses different preselection criteria to ours. Namely, Secrest et al. (2015) required all their sources to have $S/N \geq 5$ in all the first three WISE channels, while we do not use the $W3$ band for selection or classification. We note that due to AllWISE observational limitations, the 12 μm S/N requirement of Secrest et al. (2015) leads to very non-uniform sky coverage of their AGN candidates, varying over an order magnitude in surface density on different patches of the sky (Figs. 1 and 2 therein); no such issues are evident in our sample except for the Galactic equator area. On the other hand, our QSO candidate sample is much shallower than the AllWISEAGN one because of our requirements of $W1 < 16$
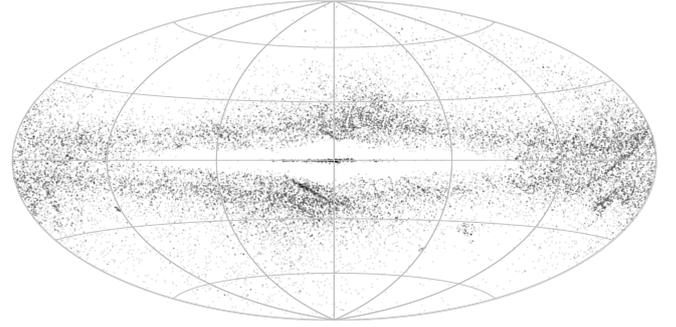
Fig. 14. Sky distribution of AllWISE anomalous sources with $0 \leq W1 - W2 < 0.8$ (26 990 objects).

and $W2 < 16$, while no such cuts were applied in Secrest et al. (2015); the latter sample thus reaches formally to the full depth of AllWISE (modulo the additional $W3$ preselection), which is $W1 \sim 17$ on most of the sky, and over $W1 = 18$ by the ecliptic poles (Jarrett et al. 2011). There are over 25 000 common sources between our QSO candidate dataset and AllWISEAGN, which means that almost 30% of our sample outside the Galaxy was not identified by Secrest et al. (2015) as AGN candidates. Taking into account that our dataset is one magnitude shallower than AllWISEAGN, we expect that applying our method at the full depth of AllWISE will bring of the order of 100 000 more QSO candidates not contained in the Secrest et al. (2015) sample and uniformly distributed over the sky. We plan to work on such a selection in the near future.

4.3. Anomalies with intermediate $W1 - W2$ colour: mixture of stars and compact galaxies

We find approximately 27 000 anomalous sources with intermediate $W1 - W2$ colours ($0 \leq W1 - W2 < 0.8$), mostly located at low Galactic latitudes but outside the Bulge area, except for a small fraction by the Galactic Centre which again are supposedly photometric artefacts (Fig. 14). Interestingly, there is an enhancement in surface density of these outliers also by the Galactic Anticentre. Only 14% of them are located at $|b_{Gal}| > 30°$, i.e. on half of the sky. Regarding their photometric properties in WISE, these sources are mostly faint, peaking at the limit of the catalogue, $W1 = 16$. About half of them have $W3$ detections, which is very different from the AGN candidate case from the previous section. Moreover, they appear to be very compact, having `w1mag13` values below 0.1. This causes their anomalous behaviour for the algorithm, as practically no sources in the training sample have this property (cf. Fig. 1).

Similarly to what was done in Sect. 4.2, we paired up this sample with external datasets. Unlike in that case, here almost exactly a half of them have counterparts in 2MASS PSC, and the sky distribution of the matches roughly follows that of this anomaly sample. Except for a handful of real artefacts from the Galactic Bulge, all these objects are faint in the JHK_s bands. On the other hand, exactly none (0) of these anomalies have a counterpart in 2MASS XSC. Altogether, this means that except for obvious artefacts, these sources are either stars or compact galaxies unresolved by 2MASS.

Further evidence that they are a mixture of these two source types comes from a cross-match with the SDSS photometric catalogue. Here we find only ~4500 matches, partly because this outlier dataset overlaps with SDSS footprint only to small
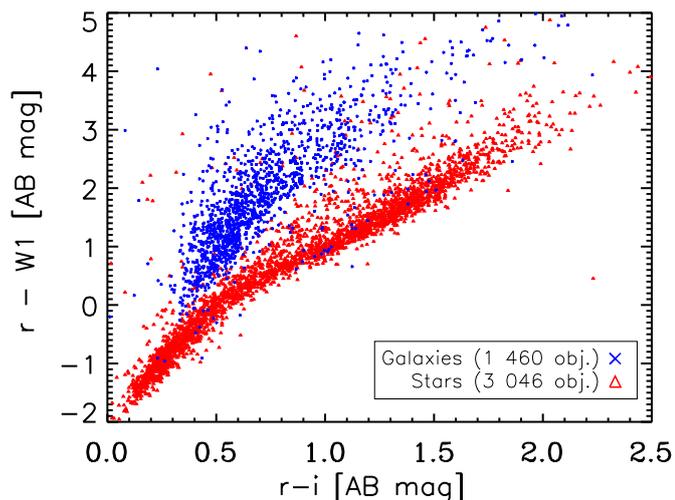
Fig. 15. Optical-infrared colour-colour diagram for sources in the All-WISE anomaly sample with intermediate $0 \leq W1 - W2 < 0.8$ and also present in the SDSS photometric dataset. Blue and red dots are respectively stars and galaxies according to SDSS morphological classification. All magnitudes are here AB; WISE $W1$ was converted following Jarrett et al. (2011).

extent, with practically no anomalies in the north Galactic cap where the SDSS coverage is the best. As shown in Prakash et al. (2015), a combination of optical ri bands and WISE $W1$ can be used for an efficient separation of stars from galaxies (see also DESI Collaboration et al. 2016 for a similar separation using the z band rather than i). We indeed observe such a division in our anomaly sample, as shown in Fig. 15; magnitudes are in the AB system for a straightforward comparison with the SDSS and DESI studies. Colour-coding by SDSS morphological classification (blue=stars; red=galaxies) confirms the two-class nature of the part of the anomaly sample which has SDSS counterparts. The galaxies from this anomaly subset matched with SDSS (1500) are also present in the SDSS DR12 photometric redshift catalogue (Beck et al. 2016) and their redshift distribution is slightly shifted towards smaller redshifts than the SDSS DR12 sample.

Based on the above considerations, we conclude that the All-WISE anomalies identified by OCSVM with $0 \leq W1 - W2 < 0.8$ colour are a mixture of stars (probably dominating), compact galaxies outside the local volume, and a handful of actual artefacts. However, only with the addition of optical photometry does the distinction between stars and galaxies become more straightforward. To distinguish between stars and galaxies without resorting to optical measurements, i.e. based on WISE data alone, proper motion measurements could possibly be used. We note that Kurcz et al. (2016) made an attempt to use proper motions as a discriminating parameter for automated source classification in AllWISE data, but no improvement was found after adding them, and actually the accuracy of the classifier was decreasing. This effect can be attributed to the fact that the WISE proper motions are not yet accurate enough to be used in source classification; they are reliable only for a small subset of WISE with high signal-to-noise ratio (Kirkpatrick et al. 2014, 2016). In future WISE data releases (Faherty et al. 2015; Meisner et al. 2017a,b) and in planned surveys like LSST, proper motions should become an important parameter to be used in classification schemes.

5. Summary and future prospects

In this work we demonstrated the power of automatic semi-supervised outlier detection, based on one-class reformulation of the SVM algorithm, and applied it to the WISE survey. By design the algorithm creates a model of standard patterns and relations in the data through training on a set of known objects. Then, data from a target set can be fitted to the model of normality and classified as either known or unknown.

The most relevant feature of the OCSVM algorithm is its ability to detect real outliers among the sources in the given dataset. In the present application to AllWISE, we found three main groups of such anomalies. The first group contains actual photometric artefacts located in areas of high surface density which have underestimated fluxes at 4.6 μm, most likely due to blending, and thus unphysical mid-IR colours. The second group includes real astrophysical objects, whose nature is consistent with that of a dusty AGN population. Their main outlying property is their very red mid-IR colour, which caused these sources to become unclassifiable in the classical approaches to automated AllWISE object divisions. The third group of anomalous sources are a mix of IR-bright stars and compact galaxies, underrepresented in the optically selected training set.

By adding these so far missing sources specific to mid-IR selection but not present in optically driven training sets, automated source separation in AllWISE data should provide more reliable results that was possible with the classical SVM approach presented in Kurcz et al. (2016). For the best performance of the SVM-based automated source classification it would be advisable to apply a novelty detection on the data before the traditional classification is conducted. This approach should ensure that the training sample includes a sufficient number of object types contained within the survey and should bring insight into how well the training sample built from another survey can represent the data to be classified.

The OCSVM algorithm can be used not only as an outlier detector, but also as a means of testing the adequacy of training samples for fully supervised classification methods. When a pattern in the data does not match any of the previously learned templates, a standard supervised classifier will assign the membership to a randomly chosen class. The versatility of the OCSVM algorithm to deal with outlying and otherwise unclassifiable data should be taken advantage of by using it as a primary step to create reliable training samples as well as to provide insight into what types of objects the supervisor should expect to find.

In a more remote future, algorithms such as OCSVM should prove essential in the efficient search for novel, unexpected, or just rare objects in the ever growing volume of data collected by planned surveys like SPICA, SKA, or LSST.

References

Agyemang, M., Barker, K., & Alhajj, R. 2006, Intell. Data Anal., 10, 521

Angiulli, F., Fassetti, F., & Palopoli, L. 2009, ACM Trans. Database Syst., 34, 7

Banerji, M., McMahon, R. G., Hewett, P. C., Gonzalez-Solares, E., & Koposov, S. E. 2013, MNRAS, 429, L55

Baron, D., & Poznanski, D. 2017, MNRAS, 465, 4530

Basu, S., Bilenko, M., & Mooney, R. J. 2004, in Proc. Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04 (New York, NY, USA: ACM), 59

Batuwita, R., & Palade, V. 2013, Class Imbalance Learning Methods for Support Vector Machines (John Wiley and Sons, Inc.), 83

Beaumont, C. N., Williams, J. P., & Goodman, A. A. 2011, ApJ, 741, 14

Beck, R., Dobos, L., Budavári, T., Szalay, A. S., & Csabai, I. 2016, MNRAS, 460, 1371

Benjamin, R. A., Churchwell, E., Babler, B. L., et al. 2003, PASP, 115, 953

Bilicki, M., Jarrett, T. H., Peacock, J. A., Cluver, M. E., & Steward, L. 2014, ApJS, 210, 9

Bilicki, M., Peacock, J. A., Jarrett, T. H., et al. 2016, ApJS, 225, 5

Blanton, M. R., & Roweis, S. 2007, AJ, 133, 734

Bolton, A. S., Schlegel, D. J., Aubourg, É., et al. 2012, AJ, 144, 144

Brandl, B. R., Bernard-Salas, J., Spoon, H. W. W., et al. 2006, ApJ, 653, 1129

Cavuoti, S., Brescia, M., D'Abrusco, R., Longo, G., & Paolillo, M. 2014, MNRAS, 437, 968

Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, ArXiv e-prints [arXiv:1612.05560]

Chandola, V., Banerjee, A., & Kumar, V. 2009, ACM Comput. Surv., 41, 15

Chapelle, O., & Zien, A. 2005, in AISTATS 2005, Max-Planck-Gesellschaft, 57

Chapelle, O., Schölkopf, B., & Zien, A. 2006, Semi-Supervised Learning, Adaptive computation and machine learning (Cambridge, USA: MIT Press), 508

Cluver, M. E., Jarrett, T. H., Hopkins, A. M., et al. 2014, ApJ, 782, 90

Cortes, C., & Vapnik, V. 1995, Mach. Learn., 20, 273

Cutri, R. M., Wright, E. L., Conrow, T., et al. 2013, Explanatory Supplement to the AllWISE Data Release Products, Tech. rep., ed. R. M. Cutri et al.

DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016, ArXiv e-prints [arXiv:1611.00036]

Donoso, E., Yan, L., Stern, D., & Assef, R. J. 2014, ApJ, 789, 44

Fadely, R., Hogg, D. W., & Willman, B. 2012, ApJ, 760, 15

Faherty, J. K., Alatalo, K., Anderson, L. D., et al. 2015, ArXiv e-prints [arXiv:1505.01923]

Hambly, N. C., MacGillivray, H. T., Read, M. A., et al. 2001, MNRAS, 326, 1279

Han, J., Kamber, M., & Pei, J. 2011, Data Mining: Concepts and Techniques, 3rd edn. (San Francisco, USA: Morgan Kaufmann Publishers Inc.)

Hautamaki, V., Karkkainen, I., & Franti, P. 2004, in Proc. Pattern Recognition, 17th International Conference on (ICPR'04) Vol. 3, ICPR '04 (Washington, DC, USA: IEEE Computer Society), 430

Hawkins, S., He, H., Williams, G. J., & Baxter, R. A. 2002, in Proc. 4th International Conference on Data Warehousing and Knowledge Discovery, DaWaK 2000 (London, UK: Springer-Verlag), 170

Heinis, S., Kumar, S., Gezari, S., et al. 2016, ApJ, 821, 86

Ho, T. K. 1998, IEEE Trans. Pattern Anal. Mach. Intell., 20, 832

Hodge, V., & Austin, J. 2004, Artif. Intell. Rev., 22, 85

Hoffmann, H. 2007, Pattern Recogn., 40, 863

Hoyle, B. 2016, Astron. Comput., 16, 34

Jarrett, T. H., Chester, T., Cutri, R., et al. 2000, AJ, 119, 2498

Jarrett, T. H., Cohen, M., Masci, F., et al. 2011, ApJ, 735, 112

Jarrett, T. H., Cluver, M. E., Magoulas, C., et al. 2017, ApJ, 836, 182

Jolliffe, I. 2002, Principal component analysis (New York: Springer Verlag)

Kirkpatrick, J. D., Schneider, A., Fajardo-Acosta, S., et al. 2014, ApJ, 783, 122

Kirkpatrick, J. D., Kellogg, K., Schneider, A. C., et al. 2016, ApJS, 224, 36

Kovács, A., & Szapudi, I. 2015, MNRAS, 448, 1305

Krakowski, T., Małek, K., Bilicki, M., et al. 2016, A&A, 596, A39

Kriegel, H.-P., Kröger, P., & Zimek, A. 2009, ACM Trans. Knowl. Discov. Data, 3, 1

Kurcz, A., Bilicki, M., Solarz, A., et al. 2016, A&A, 592, A25

Langone, R., Mall, R., Alzate, C., & Suykens, J. A. K. 2015, ArXiv e-prints [arXiv:1505.00477]

Le, T., Tran, D., Ma, W., & Sharma, D. 2010, An optimal sphere and two large margins approach for novelty detection, 2010 Int. Joint Conf. Neural Network (IJCNN)

Le, T., Tran, D., Ma, W., & Sharma, D. 2011, Multiple Distribution Data Description Learning Algorithm for Novelty Detection, Adv. Knowledge Discovery Data Mining, Proc., 246

Liu, Y.-H., Liu, Y.-C., & Chen, Y.-Z. 2011, Expert Syst. Appl., 38, 6222

Mainzer, A., Bauer, J., Cutri, R. M., et al. 2014, ApJ, 792, 30

Małek, K., Solarz, A., Pollo, A., et al. 2013, A&A, 557, A16

Manevitz, L., & Yousef, M. 2007, Neurocomput., 70, 1466

Markou, M., & Singh, S. 2003, Signal Processing, 83, 2499

Marton, G., Tóth, L. V., Paladini, R., et al. 2016, MNRAS, 458, 3479

Mateos, S., Alonso-Herrero, A., Carrera, F. J., et al. 2012, MNRAS, 426, 3271

Meisner, A. M., Lang, D., & Schlegel, D. J. 2017a, ArXiv e-prints [arXiv:1705.06746]

Meisner, A. M., Lang, D., & Schlegel, D. J. 2017b, AJ, 153, 38

Mercer, J. 1909, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 209, 415

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. 2015, e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, r package version 1.6-7

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., & Müller, K.-R. 1999, in Proceedings of the 1999 IEEE Signal Processing Society Workshop, 9, Max-Planck-Gesellschaft (IEEE), 41

Murphy, K. P. 2012, Machine Learning: A Probabilistic Perspective (The MIT Press)

Pollo, A., Rybka, P., & Takeuchi, T. T. 2010, A&A, 514, A3

Prakash, A., Licquia, T. C., Newman, J. A., & Rao, S. M. 2015, ApJ, 803, 105

R Core Team. 2013, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria

Rahman, M., Ménard, B., & Scranton, R. 2016, MNRAS, 457, 3912

Sangeetha, R., & Kalpana, B. 2010, A Comparative Study and Choice of an Appropriate Kernel for Support Vector Machines, eds. V. V. Das, & R. Vijaykumar (Berlin, Heidelberg: Springer Berlin Heidelberg), 549

Sauvage, M., Tuffs, R. J., & Popescu, C. C. 2005, Space Sci. Rev., 119, 313

Schölkopf, B., Smola, A. J., & Müller, K.-R. 1999, in Advances in Kernel Methods, ed. B. Schölkopf, C. J. C. Burges, & A. J. Smola (Cambridge, USA: MIT Press), 327

Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., & Platt, J. 2000, Adv. Neural Inf. Process. Syst., 582

SDSS Collaboration, Albareti, F. D., Allende Prieto, C., et al. 2016, ArXiv e-prints [arXiv:1608.02013]

Secrest, N. J., Dudik, R. P., Dorland, B. N., et al. 2015, ApJS, 221, 12

Shawe-Taylor, S., & Cristianini, N. 2004, Kernel Methods for Pattern Analysis (Cambridge, UK: Cambridge, UP)

Shi, F., Liu, Y.-Y., Sun, G.-L., et al. 2015, MNRAS, 453, 122

Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163

Solarz, A., Pollo, A., Takeuchi, T. T., et al. 2012, A&A, 541, A50

Solarz, A., Pollo, A., Takeuchi, T. T., et al. 2015, A&A, 582, A58

Stern, D., Assef, R. J., Benford, D. J., et al. 2012, ApJ, 753, 30

Tax, D. M., & Duin, R. P. 2004, Mach. Learn., 54, 45

Tax, D. M. J., & Duin, R. P. W. 1999, Patt. Recog. Lett., 20, 1191

Taylor, M. B. 2005, in Astronomical Data Analysis Software and Systems XIV, eds. P. Shopbell, M. Britton, & R. Ebert, ASP Conf. Ser., 347, 29

Taylor, M. B. 2006, in Astronomical Data Analysis Software and Systems XV, eds. C. Gabriel, C. Arviset, D. Ponz, & S. Enrique, ASP Conf. Ser., 351, 666

Škoda, P., Shakurova, K., Koza, J., & Palička, A. 2016, ArXiv e-prints [arXiv:1612.07549]

Škoda, P., Palička, A., Koza, J., & Shakurova, K. 2017, IAU Symp., 325, 180

Vapnik, V. N. 1995, The nature of statistical learning theory (New York, USA: Springer-Verlag New York, Inc.)

Vapnik, V., & Chervonenkis, A. 1974, Theory of Pattern Recognition [in Russian] (Moscow: Nauka), (German Translation: W. Wapnik & A. Tscherwonenkis), Theorie der Zeichenerkennung, Akademie–Verlag, Berlin, 1979

Walker, H. J., Volk, K., Wainscoat, R. J., Schwartz, D. E., & Cohen, M. 1989, AJ, 98, 2163

Wolf, C., Meisenheimer, K., Röser, H.-J., et al. 2001, A&A, 365, 681

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868

Yan, L., Donoso, E., Tsai, C.-W., et al. 2013, AJ, 145, 55

Yang, J., & Wang, W. 2003, Cluseq: efficient and effective sequence clustering

York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579

Zhang, Y., & Zhao, Y. 2004, A&A, 422, 1113