An all-sky support vector machine selection of WISE YSO candidates

G. Marton,1* L. V. Tóth,2 R. Paladini,3 M. Kun,1 S. Zahorecz,2,4 P. McGehee3 and Cs. Kiss1

1Konkoly Observatory, Research Centre for Astronomy and Earth Sciences, Hungarian Academy of Sciences, H-1121 Budapest, Hungary
2Department of Astronomy, Loránd Eötvös University, Pázmány P. s.1/a, H-1117 Budapest, Hungary
3Infrared Processing Analysis Center, California Institute of Technology, 770 South Wilson Ave., Pasadena, CA 91125, USA
4European Southern Observatory, Karl-Schwarzschild-Str. 2, D-85748 Garching bei München, Germany

ABSTRACT
We explored the AllWISE catalogue of the Wide-field Infrared Survey Explorer (WISE) mission and identified Young Stellar Object (YSO) candidates. Reliable 2MASS and WISE photometric data combined with Planck dust opacity values were used to build our data set and to find the best classification scheme. A sophisticated statistical method, the support vector machine (SVM) is used to analyse the multidimensional data space and to remove source types identified as contaminants (extragalactic sources, main-sequence stars, evolved stars and sources related to the interstellar medium). Objects listed in the SIMBAD data base are used to identify the already known sources and to train our method. A new all-sky selection of 133 980 Class I/II YSO candidates is presented. The estimated contamination was found to be well below 1 per cent based on comparison with our SIMBAD training set. We also compare our results to that of existing methods and catalogues. The SVM selection process successfully identified >90 per cent of the Class I/II YSOs based on comparison with photometric and spectroscopic YSO catalogues. Our conclusion is that by using the SVM, our classification is able to identify more known YSOs of the training sample than other methods based on colour–colour and magnitude–colour selection. The distribution of the YSO candidates well correlates with that of the Planck Galactic Cold Clumps in the Taurus–Auriga–Perseus–California region.

Key words: methods: data analysis – methods: statistical – stars: pre-main-sequence – stars: protostars – infrared: general – infrared: stars.

1 INTRODUCTION
The amount of data collected by infrared (IR) satellites and observatories has been continuously increasing over the past three decades. The evolution of the detectors allowed us to explore the interstellar medium (ISM) and embedded objects in more and more detail. IRAS (Neugebauer et al. 1984) provided ~350 000 objects with flux above 0.5 Jy at 12 µm. Recently, based on observations of the Wide-field Infrared Survey Explorer (WISE; Wright et al. 2010) IR satellite, more than 700 million sources with flux above 1 mJy were catalogued in the AllWISE Data Release (Cutri et al. 2013). IR luminous objects cover a broad spectrum of object types. Extragalactic sources, especially galaxies with ongoing star formation or active galactic nuclei (AGNs), show similar spectral energy distribution (SED) to that of the Young Stellar Objects (YSOs). Evolved stars eject dust into their outer envelopes, which has IR colours analogous to the dust surrounding YSOs in their early evolutionary stages. In this work, we identify the AllWISE sources by searching for a close counterpart in the SIMBAD data base, using a 5 arcsec radius. The closest SIMBAD source within this radius is associated with the AllWISE object. Fig. 1 illustrates the surface density of the known extragalactic sources, main-sequence (MS) stars, evolved stars and YSOs in the WISE W1–W2, W2–W3 colour–colour plane and it shows that the different object types have highly overlapping WISE colours. Linear methods would fail to separate the different object types and result in samples with high percentage of contamination, so the separation requires special attention.

The complexity of the observable properties makes the object classification a fundamental and challenging problem. However, the commonly used schemes do not take advantage of all the available information. Sources are often identified on colour–colour and colour–magnitude diagrams: Gutermuth et al. (2008) characterized the Spitzer (Werner et al. 2004) IRAC (Fazio et al. 2004) colour and magnitude properties of proto- and pre-MS stars in NGC 1333, and completed the data set with MIPS (Rieke et al. 2004) and J, H and K, 2MASS (Cutri et al. 2003; Skrutskie et al. 2006) data. This method was then extended by Gutermuth et al. (2009) to...
several star-forming clouds, located within 1 kpc of the Sun. Rebull et al. (2010) also used IRAC and MIPS colours together with 2MASS data to identify YSOs in the Taurus Molecular Cloud (TMC). They note that the method does not seem to successfully weed out all the galaxies. Harvey et al. (2007) set criteria for YSO identification on large-scale maps of star-forming regions, with a primary goal of mitigation of extragalactic contamination. This kind of large-scale classification is problematic because of the wide variation of extinction. Rebull et al. (2011) also identified YSO candidates in the TMC by using WISE photometry, as described in Koenig et al. (2012). By using far-IR (four bands between 65 and 160 µm) AKARI (Murakami et al. 2007) FIS (Kawada et al. 2007) colours and flux densities, Pollo, Rybka & Takeuchi (2010) successfully separated the sources of the AKARI Bright Source Catalogue (Yamamura et al. 2010) in low-extinction regions by classifying them as either extragalactic sources or Milky Way stars. Based on a combination of far-IR AKARI and mid-IR WISE data, Tóth et al. (2014) used quadratic discriminant analysis (QDA) to identify YSO candidates. Their comparison to the known YSOs of the SIMBAD database showed that 90 per cent of the training sample YSOs were successfully reclassified as YSO candidates, while the fraction of known contaminants remained <10 per cent.

In this paper, we built a multidimensional data set containing near-IR 2MASS and mid-IR WISE data and we apply the support vector machine (SVM) method to identify potential YSO candidates. Our goal is to create a catalogue of carefully selected YSO candidates that can be used for statistical studies and that can also provide a list of potential targets for future follow-up observations. We show that SVM (Vapnik 1995), which is a commonly used tool in pattern recognition and in multidimensional classification, is able to identify higher fraction of the training sample YSOs than the regularly used polynomial selections on colour–colour and colour–magnitude planes. We have to note that, due to the lack of spectroscopic data, our training samples are based on SIMBAD identifications. Therefore, most of our comparisons are also estimates based on SIMBAD. We identify extragalactic contaminants, Galactic field stars, Galactic evolved stars and other Galactic contamination. As a result, YSO candidates are identified and an attempt is made to separate them based on their evolutionary stages. Verification of our method and comparison to existing methods of YSO selection has been done. We also investigate, in a well-known star-forming region (i.e. the Taurus–Auriga–Perseus–California molecular cloud), the potential spatial correlation between the candidate YSOs and the Cold Clumps listed in the Planck Catalogue (Planck Collaboration 2015).

2 DATA AND METHOD

2.1 Data

To search for the YSO candidates, we used the AllWISE Data Release (Cutri et al. 2013), which is an improved version of the WISE All-Sky Data Release (Cutri et al. 2012). The AllWISE catalogue contains information on 747 634 026 sources. We used not only brightness values in all four WISE passbands (3.6, 4.6, 12 and 22 µm), but also the extended source flag (ext), which indicates whether or not the morphology of a source is consistent with the WISE point spread function, and also if the source is associated with or superimposed on a previously known extended object from the 2MASS Extended Source Catalog. We also used 2MASS J, H, Ks magnitudes, which are provided in the AllWISE catalogue, based on 2MASS Point Source Catalogue (Cutri et al. 2003) associations. Instead of the whole AllWISE catalogue, we used only those sources that matched the following criteria: (1) signal-to-noise ratio (SNR) > 3 in all four WISE bands and (2) 2MASS J, H, Ks magnitudes are available with photometric errors lower than 0.1 mag. Applying these criteria resulted in 8956 636 sources. These form our initial sample, which hereafter we call the W0 sample.

At the position of each source, we estimated the dust optical depth (τ) by using the 353 GHz R1.2 Planck dust opacity maps (Planck Collaboration XI 2014). This operation allowed us to include the effect of interstellar reddening in our analysis.

2.2 Support vector machines

For classification and pattern recognition in multidimensional data, one can use several statistical methods. We used the SVM, a class of supervised learning algorithm, developed by Vapnik (1995) as an extension to non-linear models of the generalized portrait algorithm. SVM calculates decision planes between different known classes of objects and applies the decision planes to objects of unknown classes. These unknown objects are classified based on their position in the multidimensional parameter space with respect to the separation boundaries. A more detailed description of the method can be found in Malek et al. (2013). Various statistical methods are usually among the major ingredients of professional statistical software packages. We used the R implementation of SVM in our work.

SVM is a supervised learning algorithm, therefore it needs a training set, which is used to determine the boundaries in the parameter space between the different object types. To find out the object types of our sources, we searched the SIMBAD data base for counterparts within 5 arcsec radius. If more than one object was located within this radius, then the type of the closest entry was used. More information about the SIMBAD object identification can be found in Ochsenbein & Dubois (1992). Following this strategy, we were able to identify 890 552 sources of the W0 sample. We are aware that individual source identifications in SIMBAD might not be fully reliable. However, for our purposes, we only need training samples that are statistically reliable. The number of sources found per object type are listed in the W0 column of Table A1.
3 CLASSIFICATION STEPS AND RESULTS

A multistep classification scheme was developed to remove the contaminating sources from our sample, and to identify the YSO candidates with high accuracy. The steps of our classification are shown in Fig. 2. Each step and the corresponding training samples are described in detail below. Although SVM is a very powerful tool, the necessity of a multistep classification scheme can be explained by considering the complex structure of the ISM. The interstellar reddening has an impact on the apparent colours of the sources, therefore it is important to take into account where different object types are typically found as a function of ISM column density. For this purpose, and by using the Planck dust opacity maps, we binned our sources according to the dust opacity value registered at their position in the sky.

3.1 Spurious source identification

Koenig & Leisawitz (2014) performed a careful examination of the AllWISE catalogue: they inspected, at the positions of the AllWISE sources, the higher resolution Spitzer images. In addition, in selected regions of the sky, they compared Spitzer source catalogues to AllWISE lists of sources. Their analysis led them to conclude that several AllWISE catalogued sources are spurious, and that many are likely ISM knots, or cirrus. Following their finding, our very first step was to identify and remove the spurious sources from our W0 sample.

To train the SVM, we checked the WISE W3 and W4 images of W0 positions in five different regions: the Galactic Centre, the California Molecular Cloud, the Galactic anticentre, the ρ Oph star-forming region and the Cepheus molecular complex. In these regions, we selected 680 positions where we were able to clearly identify a point source, and 664 positions where visual source identification was not possible. Examples for these real and spurious AllWISE detections are shown in Fig. C1 and Fig. C2, respectively. The training samples included, for these sources, the AllWISE catalogue information on the SNR, the reduced χ² value, the number of times the source was detected with SNR > 3 and the number of profile fits in the W3 and W4 bands.

With the help of the training sample described above, we classified the W0 sample into two classes: real and spurious sources. The misclassification rate, i.e. the rate of false positive and false negatives, was investigated as a function of the number of elements used in the training sample. Fig. 3 shows that the rate of misclassified sources does not change significantly with the number of elements used. The ratio of false negatives was 7.2 ± 1.4 per cent in the test. The lowest misclassification rate was achieved with the maximum number of elements in the training sample (5.7 per cent). The fraction of false positives was found to be 1.8 ± 0.6 per cent. With the maximum number of elements in the training sample, it was 1.7 per cent. The spurious and real sources were classified in their own class with 98.3 per cent and 94.3 per cent success rate, respectively. Applying the determined classification boundaries to
our initial sample, we classified 5366 238 sources as spurious and 3 590 398 sources as real. The latter constitutes what we refer to as real sample.

3.2 YSO identification process

The SIMBAD data base lists 235 object types. This large variety makes the training of the algorithm rather inefficient. To make the algorithm more powerful and to have statistically more robust training samples, we binned the object types based on similarities in their $J - H, H - K_s, K_s - W_1, W_1 - W_2, W_2 - W_3, W_3 - W_4$ colours and on the $\text{ext}$ (extended source) parameter. Figs D1–D4 show the distribution of the average colours and the average $\text{ext}$ values.

Three large groups of SIMBAD extragalactic objects were created. SIMBAD types belonging to the G1 group have low $H - K_s$ and $W_1 - W_2$ colours and have mostly high $\text{ext}$ values. Source types of G2 group are less extended based on the $\text{ext}$ value and have high $W_1 - W_2$ values. The remaining sources were classified as G3, and are mostly extended like the G1 objects but have higher colour indices at shorter wavelengths.

Evolved stars were also grouped in three large bins. E1 type objects have rather small colour indices and are more compact, while E2 types have high $J - K$ colours compared to the other object types. E3 objects have $W_2 - W_3$ colour higher than all the other evolved types.

Two bins of young SIMBAD objects were created: All colour indices of Y1 sources are higher than those of Y2 objects. Y1 sources also appear to be less compact because they have higher $\text{ext}$ values.

The SIMBAD type "o" (single star) was not further binned.

Finally, two groups of SIMBAD source types of ISM-related objects were defined. ISM1 sources appear to be more compact than ISM2.

The 11 subtypes are listed below, including all SIMBAD types associated with each of them:


(ii) G2: Active Galaxy Nucleus, Seyfert 2 Galaxy, Blazar, Seyfert Galaxy.

(iii) G3: Broad Absorption Line system, Gravitationally Lensed Image, Gravitationally Lensed Image of a Quasar, Seyfert 1 Galaxy, Quasar, Absorption Line system, Damped Ly-alpha Absorption Line system, Possible Quasar, BL Lac - type object.

(iv) Y1: Young Stellar Object, Variable Star of FU Ori type, Young Stellar Object Candidate.

(v) Y2: Variable Star of Orion Type, T Tau-type Star, T Tau Star Candidate.


(vii) E2: Possible Supergiant star, Possible Asymptotic Giant Branch Star, OH/IR star.

(viii) E3: Post-AGB Star, Post-AGB Star Candidate.

(ix) S: Single star.
known YSOs, of which only 117 were classified as extragalactic source (false positive).

### 3.2.2 Removal of MS stars

After removing the sources classified as extragalactic objects, our goal was to remove the MS stars from the remaining 3 213 272 sources. This galaxy-free sample contained 932 531 known MS stars. As it was done in case of the galaxy removal, we checked again the τ distribution of the MS stars and divided the sky into three regions. Those regions contained one third of the known MS stars, having τ < 1.1 × 10^{-4}, τ > 2.8 × 10^{-4} and in-between. Also (as in the previous step), the average colours and brightness values of the different subtypes were calculated to achieve a maximum possible separation between MS stars and all the other object types. Performing this task led us to use the colours $H-W4$, $K_s-W4$, $W1-W2$, $W2-W3$, $W3-W4$, the $H$ and $W4$ band magnitude values and the ext parameter.

Our results showed that we successfully removed 90.1 per cent of the known MS stars of the galaxy-free sample, by classifying 2 052 410 sources as MS star. At the same time, only 327 of the known YSOs were classified as MS star, meaning that 92.5 per cent of the known YSOs of the real sample were still kept.

### 3.2.3 Classification into evolved stars, ISM-related objects and YSO candidates

In the next step, we wanted to divide the remaining (galaxy- and MS star-free) 1 160 862 sources into three main categories: evolved stars (E1, E2 and E3 subtypes), ISM-related objects (ISM1 and ISM2 subtypes) and YSO candidates (Y1 and Y2 subtypes), with the goal to keep sources classified as Y1 or Y2. We were not able to identify regions, where one of the object types was dominant, thus cuts in the τ value were not applied. The training sample was prepared by using $J-H$, $W2-W3$, $W3-W4$ colours, the $W2$ magnitude and the ext parameter.

This classification step resulted in losing only 210 of the known YSOs, while keeping 88.8 per cent of the known YSOs in the real sample. We also removed 11 589 of the known evolved stars. Compared to the number of known evolved stars in the real sample, 27.4 per cent of them were still present at this stage. We were also able to remove 62.3 per cent of the known ISM-related objects. The resulting YSO candidate sample (sources classified as Y1 or Y2) contained 751 628 sources.

### 3.2.4 Classification into YSO evolutionary classes

The last step was to separate the Class I, II and III sources (Lada 1987) with the ultimate goal of finding reliable Class I and Class II YSO candidates. Class I and II sources have a significant IR excess that originates from the dust in their circumstellar envelopes or protoplanetary discs. The Class III sources have IR colours more similar to MS stars, but still showing IR excess. The SIMBAD data base does not list information on the actual evolutionary stage of the known YSOs, therefore our training sample was prepared based on YSO catalogues from the literature, preferably listing the evolutionary classes. We used catalogues from the following papers: Allen et al. (2012), Billot et al. (2010), Chavarría et al. (2008), Connelley, Reipurth & Tokunaga (2008), Evans et al. (2009), Gutermuth et al. (2008), Gutermuth et al. (2009), Kirk et al. (2009), Koenig, Allen & Gutermuth (2008), Megeath et al. (2012), Rebull et al. (2011), Rivera-L. et al. (2011). We note that these papers and their classification methods are based on IR data, mainly obtained with the Spitzer Space Telescope (Werner et al. 2004). We also used Connelley & Greene (2010) and Fang et al. (2009), which are catalogues of spectroscopically confirmed YSOs, and Winston et al. (2007) and Winston et al. (2010) listing YSOs with X-ray data.

Based on these catalogues, we created a training sample that contained 247 Class I, 1925 Class II and 313 Class III objects (the latter category includes sources of Koenig et al. (2008), being stars in star-forming regions, but showing mostly photospheric colours). In order to simplify the classification, where listed, we considered Transition Disc and Flat SED objects as Class II sources.

The first attempt to classify our YSO candidates into evolutionary stages failed, and the contamination caused by asymptotic giant branch star candidates (‘AB?’ as listed in SIMBAD) was still high. Therefore, the following step was additionally performed: the SVM was trained by using three subtypes, the Class I/II, the ClassIII- and the ‘AB?’ stars. Majority of the ‘AB?’ objects are those identified by Robitaille et al. (2008), and we disagree with their statement ‘YSOs and AGB stars can be mostly separated by simple colour–magnitude selection criteria’. The $J-H$, $H-W3$, $W1-W2$, $W1-W4$ and $W2-W3$ colours were used along with the $W2$ magnitude and the ext parameter, to create our training sample.

As a final result, we classified 133 980 sources as Class I/II candidates. Their surface density distribution is shown on Fig. 6. Compared to the training sample of 247 Class I, 1925 Class II and 313 Class III objects, the resulting Class I/II candidate sample contains 240 of the known Class I, 1824 of the known Class II and 79 of the known Class III sources. This means that 95 per cent of the known Class I+II sources were successfully kept, while 74.8 per cent of the Class III sources were removed. Likewise, in this last step 63.7 per cent of the known AGB star candidates were also removed. Fig. 7 illustrates the robustness of our method, as it shows the significant overlap between the known extragalactic sources, the known field stars, and our samples classified as Class I/II candidates and Class III+ candidates. Our candidate catalogues are available via the VizieR service.

### 4 DISCUSSION

#### 4.1 Reliability, false positives

We carefully investigated the possible contamination present in our Class I/II candidate sample. In our real sample, the number of sources identified with SIMBAD was 1151 956, including 5685 YSOs (sources with object types ‘Y*O’, ‘TT*’, ‘Or*’, ‘FU*’,...
Figure 6. Surface density of the Class I/II YSO candidate sources classified with SVM, shown in galactic equal-area Aitoff projection. Values were calculated in $1^\circ \times 1^\circ$ bins, and are represented on linear scale from 1 to 100. The major and best known star forming regions can be easily identified.

Figure 7. $W_1$–$W_2$ versus $K_s$–$W_4$ colour–colour diagram with our Class I/II candidate sources (orange crosses) overlaid on sources classified as Class III$^+$ (green diamonds), known MS stars from SIMBAD (magenta dots) and known SIMBAD extragalactic sources (blue triangles).

‘Y*?’ or ‘TT?’). We found only 21 568 (1.9 per cent) false-positive classifications that remained in our SVM classified Class I/II candidate sample. This means that we were able to remove 98.1 per cent of the contamination, as compared with the SIMBAD training set. The 21 568 false-positive sources were further analysed in order to learn what fraction of the contamination is coming from the different SIMBAD object types. The complete list is shown in column ‘SVM’ in Table A1. Here, we created four different categories to summarize our findings, representing groups of object types that are (i) most probably contamination, like known extragalactic objects or known evolved stars, (ii) candidate SIMBAD object types, (iii) sources that are assigned a generic object type in SIMBAD, such as IR source or star, but for which there is a non-zero probability to be instead YSOs and (iv) sources of flux that might be YSOs, or are closely related to them (e.g. maser).

(i) Obvious contamination.

(a) Extragalactic source – 101 (/110 564 – 0.1 per cent).
(b) Evolved star – 830 (/13 121 – 6.3 per cent).
(c) Other – 1955.

(ii) Possible contamination – candidate object type – 1128 (/7033 – 16 per cent). Here, we have to note that 931 of the 1128 candidate type objects are from the ‘AB?’ type. 925 of them are classified as ‘AB?’ by Robitaille et al. (2008).
(iii) Possible YSOs.

Table 1. Number of known sources in our samples of the selection process. First column indicates the name of the subtype (as defined in Section 3.2). Column 2, 3 and 4 are the $W_0$, the real and the final Class I/II candidate samples.

<table>
<thead>
<tr>
<th>Subtype</th>
<th>$W_0$</th>
<th>Real</th>
<th>Class I/II</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>148 267</td>
<td>90 840</td>
<td>65</td>
</tr>
<tr>
<td>G2</td>
<td>10 311</td>
<td>6 152</td>
<td>7</td>
</tr>
<tr>
<td>G3</td>
<td>12 729</td>
<td>8 572</td>
<td>29</td>
</tr>
<tr>
<td>E1</td>
<td>13 208</td>
<td>12 940</td>
<td>631</td>
</tr>
<tr>
<td>E2</td>
<td>3429</td>
<td>2925</td>
<td>1118</td>
</tr>
<tr>
<td>E3</td>
<td>109</td>
<td>106</td>
<td>43</td>
</tr>
<tr>
<td>ISM1</td>
<td>912</td>
<td>745</td>
<td>183</td>
</tr>
<tr>
<td>ISM2</td>
<td>515</td>
<td>319</td>
<td>173</td>
</tr>
<tr>
<td>S</td>
<td>973 629</td>
<td>932 733</td>
<td>11 990</td>
</tr>
<tr>
<td>Y1</td>
<td>9268</td>
<td>4930</td>
<td>3705</td>
</tr>
<tr>
<td>Y2</td>
<td>1128</td>
<td>755</td>
<td>637</td>
</tr>
</tbody>
</table>

(a) Single star – 11 990 (/932 531 – 1.3 per cent).
(b) IR source – 1618 (/6701 – 24.1 per cent).
(c) Sources related to ISM – 413 (/1226 – 33.7 per cent).
(d) Variable star – 1424 (/16 537 – 8.6 per cent).

(iv) Possibly related to YSOs.

(a) Radio, mm, sub-mm or X-Ray source – 185 (/3089 – 6 per cent).
(b) Maser 49 (/71 – 69 per cent).

As it is shown in the list, the total number of obvious and possible contamination is very low in the candidate sample, 4013 (0.35 per cent) sources in total. The number of obvious contaminants only, is even lower, 2886 sources (0.25 per cent).

We also made a comparison based on the binned SIMBAD subtypes defined in Section 3.2. Table 1 details the number of elements for each subtype, found in the samples indicated by the table columns. We notice that only 101 sources (0.1 per cent) of the galaxy subtypes (G1, G2 and G3) are still present in the final Class I/II sample. The remaining number of evolved stars was found to be 1792 that is 11.2 per cent of the total E1$^+$E2$^+$E3 subtypes in the real sample. We note again that 925 of the 1792 (52 per cent) are asymptotic giant branch star candidates (‘AB?’) of Robitaille et al. (2008). Without these objects, the total number of E1$^+$E2$^+$E3 sources would be only 867. The fraction of remaining single stars (as listed in SIMBAD) is only 1.3 per cent. This small fraction corresponds to 11 990 sources, which is higher than the number of SIMBAD YSOs, but it is also a very generic object type that does not prevent these sources from being YSOs in reality. The number of objects from ISM-related types (ISM1$^+$ISM2) is 356,
corresponding to 33.5 per cent of the real sample. We emphasize once again that the SIMBAD associations are rather generic, therefore some of the sources might actually turn out to be YSOs.

The Sloan Digital Sky Survey DR-9 (SDSS DR-9; Ahn et al. 2012) flags the object types for all their sources indicating whether the source is thought to be a galaxy or a star based on morphology. A cross-correlation with SDSS DR-9 allowed us a different estimation on the fraction of false-positive classifications. Of the SIMBAD YSOs in the real sample, 1029 of the 5685 known YSOs have a counterpart in SDSS DR-9 (using a searching radius of 5 arcsec). 14 per cent of them (144) were flagged as galaxy in their catalogue. We also cross-checked the known YSOs in our final Class I/II candidate sample. 814 of them were found in the SDSS of which 106 (13 per cent) are flagged as galaxy. Out of the total 133 980 sources classified as Class I/II, 5840 were found in the SDSS catalogue, 580 of them are flagged as galaxy (9.9 per cent). We conclude that our final SVM sample of candidate Class I/II does not contain a higher fraction of extragalactic contaminants than the sample of known YSOs in SIMBAD.

4.2 Completeness, false negatives

In the process of identifying the Class I/II sources candidates, a number of known YSOs were lost by either misclassification (false negatives) or because they were not detected by WISE. The completeness of our sample was investigated in three different ways. (i) First, we searched our W0, real and final Class I/II sample for the known YSOs in SIMBAD; (ii) then we looked, in the same samples as in (i), for YSOs listed in public photometric catalogues (see Section 3.2.4) (iii) finally, again using the samples as in (i), we looked for spectroscopically confirmed YSOs.

4.2.1 Comparison to SIMBAD YSOs

First, we considered all the known SIMBAD YSOs and searched for them in our W0, real and final Class I/II samples. The total number of known YSOs in the SIMBAD data base was 46 453 at the time of our investigation (21 186 ‘Y*O’, 20 716 ‘Y*?’, 4 496, 4 039, 733, 61, 20 and 1047, respectively) correlating with SIMBAD, we found a total number of 10 396 ‘TT*’, 237 ‘TT?’, 33 ‘FU*’ and 2450 ‘Or*’). After cross-checking the known YSOs in our final Class I/II candidate sample. 814 of them were found in the SDSS of which 106 (13 per cent) are flagged as galaxy. Out of the total 133 980 sources classified as Class I/II, 5840 were found in the SDSS catalogue, 580 of them are flagged as galaxy (9.9 per cent). We conclude that our final SVM sample of candidate Class I/II does not contain a higher fraction of extragalactic contaminants than the sample of known YSOs in SIMBAD.

<table>
<thead>
<tr>
<th>Region</th>
<th>Paper</th>
<th>Class I/II</th>
<th>Class III</th>
<th>W0</th>
<th>Class I/II</th>
<th>Class III</th>
<th>Real</th>
<th>Class I/II</th>
<th>Class III</th>
<th>SVM Class I/II</th>
<th>Class III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allen et al. (2012)</td>
<td>Cepheus OB3</td>
<td>1135</td>
<td>1440</td>
<td>435</td>
<td>279</td>
<td>209</td>
<td>29</td>
<td>198</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Billot et al. (2010)</td>
<td>Vul OB1</td>
<td>703</td>
<td>153</td>
<td>259</td>
<td>82</td>
<td>160</td>
<td>54</td>
<td>134</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chavarria et al. (2008)</td>
<td>S254–S258</td>
<td>252</td>
<td>210</td>
<td>63</td>
<td>42</td>
<td>18</td>
<td>10</td>
<td>16</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evans et al. (2009)</td>
<td>–</td>
<td>942</td>
<td>79</td>
<td>294</td>
<td>37</td>
<td>214</td>
<td>28</td>
<td>204</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gutermuth et al. (2008)</td>
<td>NGC 1333</td>
<td>133</td>
<td>–</td>
<td>57</td>
<td>–</td>
<td>46</td>
<td>–</td>
<td>43</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kirk et al. (2009)</td>
<td>Cepheus flare</td>
<td>128</td>
<td>13</td>
<td>97</td>
<td>10</td>
<td>76</td>
<td>5</td>
<td>69</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Megeath et al. (2012)</td>
<td>Orion A and B</td>
<td>2284</td>
<td>329</td>
<td>1023</td>
<td>168</td>
<td>631</td>
<td>69</td>
<td>603</td>
<td>53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rebull et al. (2011)</td>
<td>NAN complex</td>
<td>1149</td>
<td>112</td>
<td>498</td>
<td>98</td>
<td>264</td>
<td>84</td>
<td>217</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winston et al. (2007)</td>
<td>Serpens cloud core</td>
<td>115</td>
<td>22</td>
<td>46</td>
<td>2</td>
<td>31</td>
<td>2</td>
<td>22</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winston et al. (2010)</td>
<td>NGC 1333</td>
<td>54</td>
<td>41</td>
<td>36</td>
<td>11</td>
<td>34</td>
<td>6</td>
<td>31</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>11 229</td>
<td>2399</td>
<td>4097</td>
<td>729</td>
<td>2336</td>
<td>287</td>
<td>2154</td>
<td>79</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Recently, Koenig & Leisawitz (2014, hereafter KL14) published a method to reduce the number of spurious sources and to identify YSOs in the WISE bands. An additional 249 sources (14.4 per cent) turned out to be spurious detections, while 434 of the remaining 475 (91.3 per cent) were successfully re-classified.

We also investigated whether the KL14 method is more sensitive to Class II YSOs rather than to more evolved class types. Using the same sources that we used in Section 3.2.4, we found that KL14 method finds 95 of the 247 Class I sources (SVM finds 240), 922 of the 1925 Class II sources (1824 with SVM) and 47 of the 313 Class III objects (SVM result is 79). The combined results show that SVM is able to retrieve 2064 of the 2172 Class I/II sources (95 per cent), while the KL14 method finds only 589 (27 per cent). These results suggest that the overlap between the two methods is rather small; however, the results based on the SIMBAD catalogue are more reliable.

We also investigated the BLG method. The comparison between the two methods is provided in Table A1.

### 4.4 Comparison with the QDA

In our previous work (Tóth et al. 2014), the QDA (McLachlan 1992) technique was used to identify YSO candidates using far-IR AKARI and mid-IR WISE data. In this section, we compare the KL14 method to the initial W0 sample. We can see that the number of false-positive extragalactic objects in our Class I/II selection is 101 (0.05 per cent compared to the real sample), while it is 5889 (3.4 per cent) in the KL14 sample. The fraction of known SIMBAD YSOs that were recovered with our method is higher (41.7 per cent) than with the KL14 one (32.1 per cent). On the other hand, the KL14 method successfully eliminated 95.9 per cent of the evolved stars (89.3 per cent with the SVM).

We conclude that KL14 method is very efficient in identifying and removing the Galactic contamination, but allows us to retrieve a lower number of known YSOs. Our method is more successful in removing extragalactic contamination, and it misclassifies known YSOs in a smaller fraction. A comparison of our selection to existing spectroscopic YSO catalogues is provided in Table A1.

### Table 4. Number of sources of our binned SIMBAD subtypes (first column) in the initial W0 sample (second column), in our SVM classified Class I/II candidate sample (third column) and in the YSO sample of the KL14 method (last column).

<table>
<thead>
<tr>
<th>Subtype</th>
<th>W0</th>
<th>SVM Class I/II</th>
<th>KL14 YSOs</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>148267</td>
<td>65</td>
<td>1398</td>
</tr>
<tr>
<td>G2</td>
<td>10311</td>
<td>7</td>
<td>1188</td>
</tr>
<tr>
<td>G3</td>
<td>12729</td>
<td>29</td>
<td>3303</td>
</tr>
<tr>
<td>E1</td>
<td>13208</td>
<td>631</td>
<td>58</td>
</tr>
<tr>
<td>E2</td>
<td>3429</td>
<td>1118</td>
<td>613</td>
</tr>
<tr>
<td>E3</td>
<td>109</td>
<td>43</td>
<td>16</td>
</tr>
<tr>
<td>ISM1</td>
<td>912</td>
<td>183</td>
<td>90</td>
</tr>
<tr>
<td>ISM2</td>
<td>515</td>
<td>173</td>
<td>50</td>
</tr>
<tr>
<td>S</td>
<td>973629</td>
<td>11990</td>
<td>1043</td>
</tr>
<tr>
<td>Y1</td>
<td>9268</td>
<td>3705</td>
<td>2695</td>
</tr>
<tr>
<td>Y2</td>
<td>1128</td>
<td>637</td>
<td>650</td>
</tr>
</tbody>
</table>

i.e. a large proportion of YSOs do not have reliable fluxes in all WISE and/or 2MASS bands. An additional 249 sources (14.4 per cent) turned out to be spurious WISE detections, while 434 of the remaining 475 (91.3 per cent) were successfully re-classified.

### 4.4 Comparison with the QDA

In our previous work (Tóth et al. 2014), the QDA (McLachlan 1992) technique was used to identify YSO candidates using far-IR AKARI and mid-IR WISE data. In this section, we compare the...
currently used SVM method and the QDA by repeating the steps of the analysis described above, and using the same training samples, to find out which one suits the problem better. The main difference between the two methods is their approach to the decision boundaries. Discriminant analysis techniques perform dimensionality reduction, and project the data into a subspace where they maximize the separation. On the contrary, the SVM maps the data into a higher dimensional space and a hyperplane is calculated that provides the best separation of the classes.

As a first step, the classification of real and spurious sources was repeated. QDA successfully classified as such only 77.9 per cent of the spurious sources (compared to 98.3 per cent for SVM), and only 37.9 per cent of the real sources (compared to 94.2 per cent for SVM). In this case, SVM clearly outperforms the QDA method.

As a second step, we repeated the removal of the extragalactic sources. 99.3 per cent of the sources in the training sample were successfully re-classified as extragalactic source and only 0.7 per cent (725) were misclassified by QDA. This is three times more than with SVM, for which only 237 sources were misclassified. By using SVM, we were not able to recover 2.1 per cent of the known SIMBAD YSOs, while we lost 2.3 per cent of them with QDA.

In the third step, we repeated the removal of field stars. With QDA, 17.3 per cent of the SIMBAD single stars remained in our sample, while with SVM only 8.9 per cent. Also, with SVM 5.5 per cent of the YSOs were lost while applying QDA to the same training sample resulted in the loss of 7.8 per cent of them.

As a last step of the QDA and SVM comparison, we repeated the classification of the remaining sources into three main object types, YSO candidates, evolved stars and ISM-related objects. By using SVM, we successfully re-classified 96 per cent of the known YSOs as YSO candidate. Using QDA, the success rate was only 89 per cent. The contamination caused by the remaining evolved stars is also higher with QDA. The number of evolved objects was found to be 6194 while it was 4382 with SVM.

This comparison clearly shows that while QDA and SVM are comparable in some of the steps, the overall performance of SVM is better than QDA, and it is more efficient for classifying Class I/II YSOs.

4.5 Correlation with the PGCCs

The Planck Catalogue of Galactic Cold Clumps (PGCC; Planck Collaboration 2015) is an all-sky catalogue of Galactic cold clump candidates detected by Planck. The PGCC catalogue contains 13 188 Galactic sources spread across the whole sky with a median temperature between 13 and 14.5 K and their size is described with the major and minor full width at half-maximum (FWHM) of a fitted elliptical Gaussian. Cold clumps represent the early stages of star formation (McKee & Ostriker 2007) and a spatial correlation between the cold clump and the YSO distribution is therefore expected.

To further test the robustness of our YSO classification, we analysed their position relative to the PGCCs in the Taurus–Auriga–Perseus–California molecular complex (150  <  l  <  180, –25 < b < –1), which is a well-known star-forming region. As a function of the major FWHM, we calculated the surface density of our Class I/II candidates and of the objects that we classified as MS stars. As seen on Fig. 8, the surface density of the Class I/II candidates is highest close to the PGCCs and then rapidly decreases, while that of the MS star candidates is independent of the distance measured from the cold clumps. This result strongly suggests that our candidate Class I/II sources are indeed related to the Planck cold clump population.

5 SUMMARY

The AllWISE catalogue was investigated to identify potential YSO candidates. A subset of the catalogue was used with S/N > 3 and available 2MASS J, H, K, data with photometric errors <0.1. We applied the SVM method to the initial data set of 2956 636 sources in a multistep process. Different combinations of colours and magnitudes were used, in combination with the extended source flag, in order to generate a multidimensional training samples and to remove contaminating sources. Sources of known Galactic and extragalactic types were identified with the help of the SIMBAD data base, using a 5 arcsec radius to match the AllWISE sources. As many as 133 980 objects were classified as YSO Class I/II candidates. The contamination of sources with well-known object types is <1 per cent, in comparison with our SIMBAD training set. We also compared our method to that described in KL14 and to the results obtained with a different approach, the QDA. We found that SVM outperforms the KL14 method in preserving the known YSOs and in identifying the extragalactic contamination and it is more effective than QDA.

A positional correlation analysis with the PGCC sources was performed in the case of the Taurus–Auriga–Perseus–California regions. Our Class I/II candidates appear to be characterized by a higher surface density in the proximity of the cold clumps while the MS star candidates had a uniform surface density in the field.

ACKNOWLEDGEMENTS

We thank our anonymous referee for all the useful comments that helped us to improve the manuscript. This publication makes use of data products from the WISE, which is a joint project of the University of California, Los Angeles and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National...
Aeronautics and Space Administration. This research has made use of the SIMBAD data base, operated at CDS, Strasbourg, France. This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France. This research was supported by the OTKA grants NN 111016 and K101393.

REFERENCES

Cutri R. M. et al., 2012, VizieR On-line Data Catalog: II/311
Cutri R. M. et al., 2013, VizieR On-line Data Catalog: II/311
Fazio G. et al., 2004, ApJS, 154, 10
Kawada M. et al., 2007, PASJ, 59, 389
Megeath S. T. et al., 2012, AJ, 144, 192
Murakami H. et al., 2007, PASJ, 59, 369
Ochsenbein F., Dubois P., 1992, ESO Conf. Workshop Proc. No. 43, Astronomy from Large Databases, II. European Southern Observatory, Garching, p. 405
Skrutskie M. F. et al., 2006, AJ, 131, 1163
Tóth L. V. et al., 2014, PASJ, 66, 17
Winston E. et al., 2010, AJ, 140, 266
Wright E. L. et al., 2010, AJ, 140, 1868

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Table A1. Comparison of SIMBAD sources found in our initial W0 sample, in the SVM selected Class I/II YSO sample and in the YSO candidate sample identified by the KL14 method.

Table B1. Average colours and magnitudes of SIMBAD object types.

Table B2. Average colours and magnitudes of SIMBAD object types.

Table B3. Average colours and magnitudes of SIMBAD object types.

Table B4. Average colours and magnitudes of SIMBAD object types.

Table B5. Average colours and magnitudes of SIMBAD object types.

Table B6. Average colours and magnitudes of SIMBAD object types.

Table B7. Average colours and magnitudes of SIMBAD object types.

Table B8. Average colours and magnitudes of SIMBAD object types.

Table B9. Average colours and magnitudes of SIMBAD object types.

Figure C1. Example of sources used as real for the spurious source identification training sample.

Figure C2. Example of sources selected as spurious for the spurious source identification training sample.

Figure D1. Three main groups of SIMBAD extragalactic objects.

Figure D2. Average colour indices and ext values of three main groups of SIMBAD source types of evolved objects.

Figure D3. Average colour indices and ext values of two groups of SIMBAD source types of young objects.

Figure D4. Average colour indices and ext values of two groups of SIMBAD source types of ISM-related objects.

Figure E1. Fraction of known YSOs in the W0 (black solid line), real (red dotted line), Class I/II (blue dashed line) and KL14 (green-dash–dotted line) samples as a function of brightness in the different WISE bands.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

This paper has been typeset from a TeX/LaTeX file prepared by the author.