DOI: http://dx.doi.org/10.18524/1810-4215.2017.30.114465

SUPERVISED AUTOMATIC IDENTIFICATION OF EXTRAGALACTIC SOURCES IN THE WISE×SUPERCOSMOS CATALOGUE

V. Khramtsov[1,2],V. Akhmetov[1,2]

[1]Institute of Astronomy of V.N. Karazin Kharkiv National University, Kharkiv, Ukraine

[2]Department of Astronomy and Space Informatics of V.N. Karazin Kharkiv National University, Kharkiv, Ukraine

vld.khramtsov@gmail.com, akhmetovvs@gmail.com

ABSTRACT. We present new catalogue of ~8,500,000 extragalactic objects as a result of automatic classification of WISE and SuperCOSMOS (SCOS) cross-identification product. The main goal is to create a set of candidates in extragalactic objects due to colour (photometric) features through machine learning techniques. Extragalactic sources were separated from stars in high-dimensional colour space using Support Vector Machine (SVM) classifier.

Construction of catalogue of the extragalactic objects is based on the four important procedures: 1. Cross-identification of the WISExSCOS catalogues. 2. Training set creation (Gaia DR1 and 2MASX\XSC data). 3. Feature engineering and colour-space constructing for further learning and classification. 4. Fine-tuning of SVM and separation and classification processes.

In result we got high-accuracy (~98%) algorithm for extragalactic source identification in built colour space. Product of algorithm realization is presented as photometric catalogue of the extragalactic objects and can be used for further astronomical investigations.

Keywords: Catalogues: statistics, extragalactic objects; Methods: machine learning, data analysis.

1. Introduction

Modern astronomical surveys include billions of objects, and with time amount of observed sources will increase. But not of all wide-field catalogues include spectroscopic information that allows us to study in detail these objects; at the same time, such catalogues almost always include photometric data. Therefore, the design of new algorithms to process large photometric data sets is an urgent task.

Particular problem of the catalogue analysis is a separation of extragalactic sources from the galactic ones. Without utilization of spectroscopic information, this problem usually can be solved by the most trivial way – with using morphological features (Peacock et al., 2016; Skrutskie et al., 2006); it is assumed that extragalactic sources, more often, are extended. Such approach has not to be applied successful to the deep-imagine catalogues because extragalactic sources (as active galactic nucleus (AGNs), quasars, faint distant galaxies, quasi-stellar objects (QSO) etc.) are unresolved or point-like – with photometric structure, similar to stars. Modern approach to solve this problem is separation of objects with high-dimensional diagrams representing different features (proper motion, colour, magnitude and others).

In current research we used high-dimensional colour space as feature space for star-galaxy separation. That was done from the reason that different types of objects, according to the shape of their spectral energy distribution, are in the different regions of colour diagrams. According to the fact, that galaxies are redder than stars (Pollo et al., 2010), we count on the infrared photometric information to separate stars and extragalactic sources. We choose photometric all-sky catalogues WISE and SCOS, cross-identification of which guarantees resolving sources up to $B < 20\,mag$ in the broad optical-

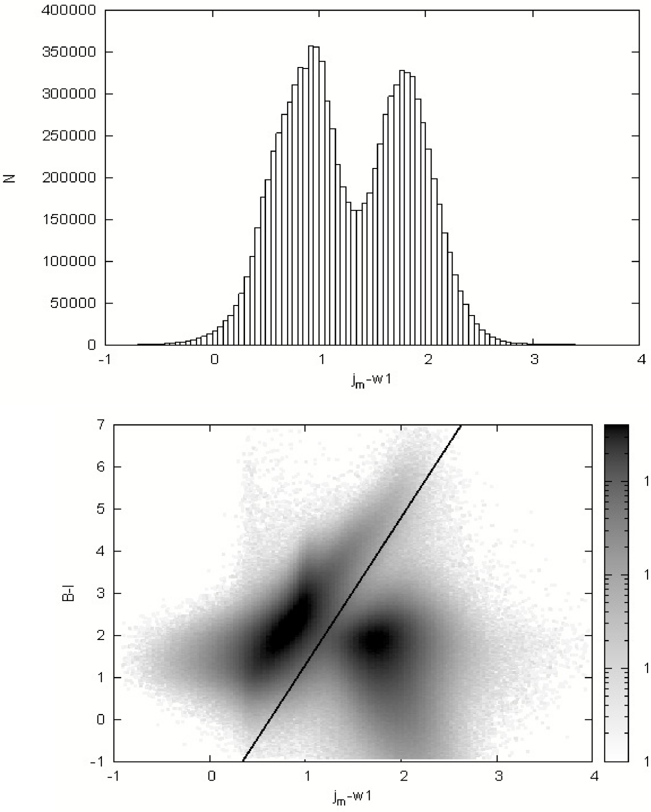

Figure 1: Colour histogram (J− W1) for PPMXL (up) and colour diagram (J− W1, B− I) for PMA (bottom). Objects in the right regions of the diagrams ($J - W1 > 1.3$) are extragalactic ones.

infrared range ($W1, W2, W3, W4, J, H, K$ (Vega), B, R, I (AB), where AB and Vega are photometry systems) and contains hundreds of millions of objects.

Method of using colour diagrams for the separation of objects into extragalactic and galactic sources was successfully applied in the creation of proper motion catalogues PPMXL (Vickers et al., 2016) and PMA (Akhmetov et al., 2017), where autors proposed to separation one- and two-colour diagrams respectively (Fig. 1).

Automatic object separation with using photometric data from the WISE and SCOS catalogues was successful carried out by several groups: (Solarz et al. 2017, in press), (Bilicki et al., 2016), (Kurcz et al., 2016), (Krakowski et al., 2016), (Kurcz et al., 2015).

The paper presents a result of classification of objects in high-dimensional colour space into two classes: candidates in extragalactic objects and galactic objects, with using linear separating classifier SVM (Vapnik, 1979). The paper is laid out as follows: the SVM method is describing in Sect.2; Sect.3 explains the procedure of data preprocessing; in Sect.4 we describe catalogue creation process (learning and classification).

2. Support Vector Machine (SVM)

For separation into classes we used Support Vector Machine (SVM) – linear supervised algorithm of classification based on building linear separating function between objects of different classes.

Classifier accepts a training set – sample of objects where each one is associated with certain feature vector and predetermined class. During learning process, the classification problem with SVM brings to finding equation of linear function that separating objects into the classes. Such function may be scalar (in 1-dimensional feature space), linear function (in 2-dimensional), plane (3-dimensional) or hyperplane (n-dimensional). After learning, unlabeled (without known class) object with the known feature vector enters into feature space and the class of this objects determines by its position relative to the separating function.

In our case, the classification problem of objects with adjusted colour indexes is a binary one. We define the labels $d = -1$ and $d = +1$ for objects, which are galactic and extragalactic objects respectively, and also, for convenience, we normalize all colour indexes $\overline{x}_i$ to the [0;1] interval. If stars and extragalactic objects are linear divided in colour space, it is obviously to select separating hyperplane, that is defined by equation: $\overline{w}^T \overline{x} + b = 0$, where $\overline{w}$ - is a normal to the hyperplane (so-called weight vector), b - is a bias. We can solve the binary classification problem using, actually, set of hyperplanes; those hyperplanes do not cross any sample of objects of different classes and, geometrically, locates between samples of stars and extragalactic objects. Width of the stripe, consisting set of hyperplanes, equals $\frac{2}{\|\overline{w}\|}$.

But it is possible to select optimal separating hyperplane, which provides equidistance to samples. We accept that optimality is observed if distance between two boundary hyperplanes (tangent to samples of stars and extragalactic sources respectively) from the set of solves, is a maximal one subject to non-existence of objects from training data between boundary hyperplanes in colour space. We can formulate conditions of optimality by follows:

$$d_i(\overline{w}\cdot \overline{\mathrm{x}}_i + \mathrm{b}) = 1$$

$$\Phi(\overline{w}) = \frac{1}{2}\|\overline{w}\|^2 \to \min$$

This optimization problem can be formulated in the Lagrangian:

$$J(\overline{w}, b, \alpha) = \Phi(\overline{w}) - \sum_i \alpha_i (\mathrm{d}_i(\overline{w}\cdot\overline{x} + \mathrm{b}) - 1)$$

where α_i – Lagrange multipliers for each object from training sample. Setting $\frac{\partial J}{\partial \overline{w}}, \frac{\partial J}{\partial b} = 0$, we obtain the solution:

$$\overline{w} = \sum_i \alpha_i d_i \overline{x}_i$$

$$\sum_i \alpha_i d_i = 0$$

allowing us to redetermine the optimization problem in the simpler, non-quadratic, form:

$$J(\alpha) = \sum_i \alpha_i - \sum_{i,j} \alpha_i \alpha_j d_i d_j \overline{x}_i \overline{x}_j$$

The weight vector $\overline{w}$ is a linear combination of the vectors from training sample, but summing occurs only for the objects for which $\alpha_i \neq 0$. This property is called sparsity and it is the main otherness of SVM from linear classifiers. Objects, for which $\alpha_i \neq 0$, are called support vectors.

In the case of linear indivisibility of samples, it is necessary to provide coordinate transformation to other one (Mercer, 1909), in which we can 1) provide search of linear separating hyperplane or 2) minimize the degree of blending of two samples. We can determine a parameter C for regulation of ratio of the minimization a classifying error and the maximization width separating stripe.

3. Data preprocessing

WISExSCOS cross-identification. Examined catalogue has been derived from the combination of two all-sky catalogues: WISE and SuperCOSMOS (SCOS). Considering the precision of determining position of sources — $< 0.15''$ for WISE (Wright et.al., 2010) and $< 0.30''$ for SCOS (Hambly et.al., 2010a), the catalogues were paired using a matching radius of 0.5". In the resulting cross-matched sample were held removing objects for which not all photometric information in bands $W1, W2, W3, W4, J, H, K$ (WISE), B, R, I (SCOS) is available. Wherein all 4 photometric bands from SCOS catalogue ($B, R1, R2, I$ (Hambly et.al., 2010b)) were used, but during filtration the averaging of $R1$ and $R2$ magnitudes for objects were done. After that, the resulting catalogue consisted of 235,232,381 objects.

Akhmetov et.al. (2017) proposed that interstellar extinction is an important application in separating plane building; to reduce false classifications, caused by interstellar reddening as distortion of colour space (Meingast et.al, 2017), we limited our WISExSCOS sample at galactic latitude $|b| > 7°$.

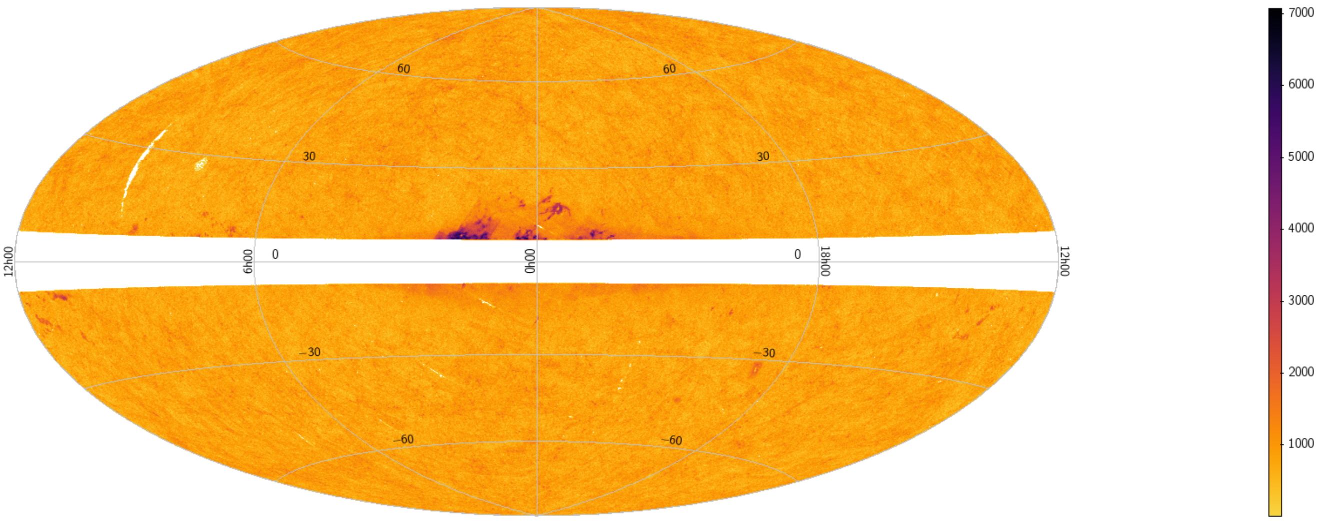

Figure 2: Aitoff density map in Galactic coordinates of 8 million objects, classified as extragalactic from WISExSCOS catalogue ($l = 0°, b = 0°$ at the centre)

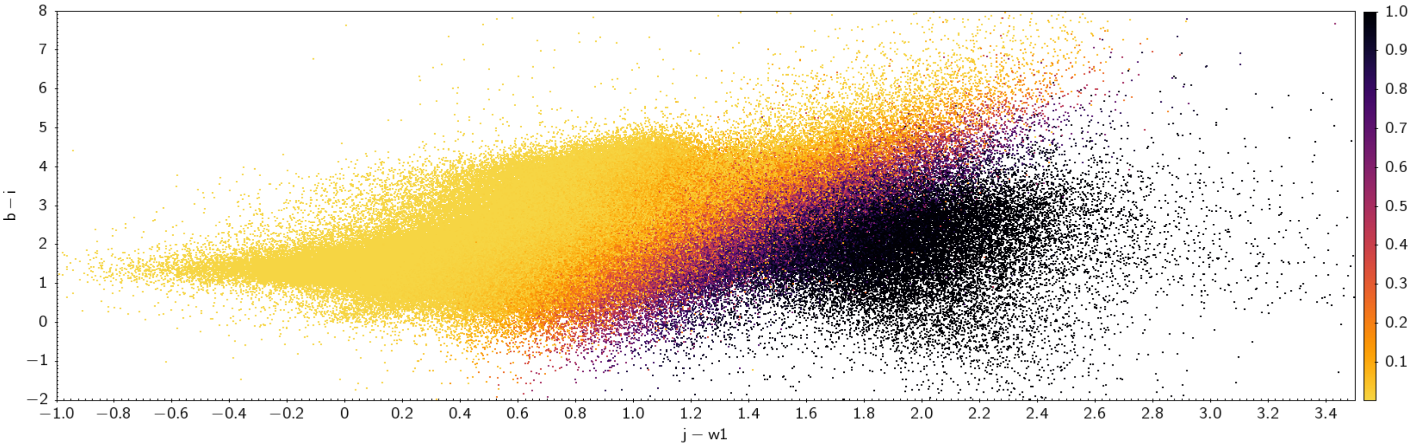

Figure 3: Colour-colour diagram (J− W1, B− I) of objects from training sample. Color bar representing the probability that object is extragalactic source.

Training set formation. Training sample was derived from 2MASX (Skrutskie et al., 2006) and GAIA DR1 (Gaia Collaboration, 2016) catalogues, assuming, that selected samples consists of the extragalactic and galactic objects respectively. We randomly extracted 50,000 objects, paired with WISExSCOS, for each catalogue. Resulting training sample included photometric information about 100,000 objects (labeled as star or extragalactic source) in 10 bands.

Colour space construction. We formed 45 colour indexes as pairwise differences of 10 magnitudes for each object from the training sample and from WISExSCOS catalogue. For building colour space for learning and classification, we produced visual feature selection of the colours of training sample as checking all two-colour diagrams.

In result we selected 6 colours, which provides clear separation of stars and extragalactic sources: $J - W1, B - I, \mathrm{W}1 - W2, \mathrm{W}2 - W3, J - W3, H - K$.

In constructed 6-dimensional colour space further learning within the training sample and classification of objects from WISExSCOS were performed. According to the selected colours it is seen, that lead role in separating extragalactic sources from stars belongs to infrared colours (Jarrett et al., 1998; Pollo et al., 2010).

4. Classification\separation process

To create the catalogue of candidates in extragalactic objects, we need to enter WISExSCOS objects alternately in the constructed colour space with existing separation hyperplane in it; the class of object determines by dint of position of the one in colour space relatively hyperplane, built during learning of SVM classifier.

Classifier were trained within 100,000 labeled objects (Gaia DR1 and 2MASX); learning was held by entering training sample and form of non-linear transformation (some kernel function). We selected Radial Basis Function (RBF) as kernel of space transformation: $\exp[-\gamma \| \overline{\mathrm{x}}_i - \overline{\mathrm{x}}_j \|^2]$, where $\| \overline{\mathrm{x}}_i - \overline{\mathrm{x}}_j \|$ - is an Euclidean distance in unreduced feature space.

Quality of learning, and also selecting parameters γ and C, were provided by classical methods of analysis of classification process – by the accuracy metric (accuracy score) using n-fold cross validation method (Kohavi, 1995).

Accuracy metric – is a way to quantify a proportion of the true classifications of the algorithm during training process:

$$acc_i = \frac{TS + TG}{TS + TG + FS + FG}$$

for i iteration, where the components of the equation are True Stars (TS), True Galaxies (TG) from the training sample, classified as stars and extragalactic objects respectively; False Stars (FS), which are real extragalactic objects from training sample, but classified as stars, and False Galaxies (FG) – labeled objects as `stars` in training data, but classified as extragalactic objects. Using n-fold cross validation, we used $n = 3$ meaning, that whole training data were split into three equal parts, and for one iteration classifier was trained on two of them; total accuracy of three-iteration learning process is defined as:

$$A_{total} = \frac{\sum_i acc_i}{3}$$

Classifier was trained with accuracy 97.9% within $C - \gamma$ combinated as $\gamma = 0.201$ and $C = 21$. For learning and classification, we used LIBSVM (Chang & Lin, 2011), integrated package for Support Vector Machine classification. Also we used Anaconda (Python distribution) – free open-source environment for large-scale data processing.

In result, we got catalogue of 8,290,477 objects (Fig. 2), identifying by SVM as extragalactic sources with probability $p > 0.5$ (Fig. 3).

5. Conclusion

Resulting photometric catalogue of candidates in extragalactic sources is one of the most accurate (in the sense of learning accuracy) catalogues of galaxies, that was got by automatic classification within colour indexes. Here we presented high quality of solving star-galaxy separation in analysis of modern deep-imaging all-sky photometric catalogues by automatic classification algorithm. However, the results obtained by us contradict the one presented earlier. Krakowski et al. (2016) showed, that final separation of galaxies from WISExSCOS catalogue consists ~15,000,000 objects; Bilicki et al. (2016) presented similar approach to identify ~19,000,000 extragalactic objects. Also, in these works the masking of high-extinction sky regions were provided much better than in this paper. Summing results, it is obvioulsy that portion of extragalactic objects in earlier investigations equals ~35% of all WISExSCOS objects; we got amount of extragalactic sources equaling ~4% of all objects. This fact can be explained by differences of training data and constructed colour space, so now a comparative analysis of the previously proposed methods with the method described in this paper is carried out.

In future we plan to focus on identification of specific extragalactic objects (quasars, AGNs, LRGs, ETGs, etc.) within our resulting catalogue of extragalactic sources.

References

Bilicki, M., Peacock, J., Jarrett, T., et al.: 2016, *ApJS*, **225**, 1, 5.

Chang, C.-C., Lin, C.-J.: 2011, *ACM Transactions on Intelligent Systems and Technology*, **2**, 27:1-27:27, software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

Gaia Collaboration: 2016, *A&A*, **592**, A2.

Hambly, N., Davenhall, A., Irwin, M., et al.: 2001a, *MNRAS*, **326**, 1315.

Hambly, N.; Irwin, M.; MacGillivray, H.: 2001b, *MNRAS*, **326**, 1295.

Jarrett, T., Chester, T., Huchra, J., et al.: 1998, *AAS*, **192**, 5515J.

Kohavi, R.: 1995, *Proceedings of the Fourteenth IJCAI*, **2** **(12)**, 1137-1143.

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

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

Kurcz, A., Krupa, M., Bilicki, M. et al.: 2015, *preprint (arXiv:1512.03604).*

Meingast, S., Lombardi, M., Alves, J.: 2017, *A&A*, **601**, A137.

Peacock, J., Hambly N., Bilicki M., et al.: 2016, *MNRAS*, **462**, 2085.

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

Skrutskie, M., Cutri, R., Stiening, R., et al.: 2006, *AJ*, **131**, 1163-1183.

Solarz, A., Bilicki, M., Gromadzki, M. et al.: 2017, *A&A* [in press].

Mercer , J.: 1909, Philos. Trans. Roy. Soc. London, A, 209, 415-446.

Vapnik, V.: 1979, Estimation of Dependences Based on Empirical Data [in Russian], Nauka, USSR.

Vickers, J., Röser, S., Grebel, E.: 2016, *AJ*, **151**, 99.

Wright, E., Eisenhardt, P., Mainzer, A., et al.: 2010, *AJ*, **140**, 1868.