CANDIDATE ASTRONOMY WITH DATA MINING: DEPROJECTING THE DISTRIBUTION OF QUASARS

OMAR LAURINO◇‡, RAFFAELE D'ABRUSCO†, GIUSEPPE LONGO‡♮, AND MASSIMO BRESCIA♮‡

Abstract. The overwhelming amount of data produced by the new generation of digital astronomical instruments observing large regions of the sky and measuring many parameters (from accurate photometry in several filters to optical spectroscopy) for hundreds of millions of extragalactic objects, has drastically changed the classical approach to astronomical research and brought, de facto, astronomy among the group of data-driven disciplines. This abrupt change in perspective has also revolutionized the framework of how astronomical research is conducted and has spurred the design and development of specific statistical models, IT tools and technological infrastructures designed ad hoc for the astronomical community to address some interesting and challenging data-related problems in their field [3]. In this paper we present a consistent application of several data mining algorithms to the astronomical realm, showing how data mining can lead to an innovative approach to the extraction of knowledge from the huge wealth of data now available for quasars: data mining allows a paradigm shift from the classical astronomy to the new "candidate astronomy". This work was carried out inside the DAME (DAta Mining and Exploration) framework and collaboration, which we will briefly introduce.

1. Introduction

The ever growing amount of astronomical data provided by the new large scale digital surveys spanning all the EM spectrum has been challenging the way astronomers carry out their everyday analysis and interpretation of astronomical information. Since the human ability to directly visualize and correlate astronomical data has been pushed to its limits, a new paradigm shift is emerging for the extraction of knowledge: data mining techniques can effectively tackle the problem of knowledge discovery and usher astronomy in new fields which could not be explored with classical techniques. We present in this paper our original approach to the problem of the characterization of the spatial distribution of quasars in the Universe, based on the application of several distinct data mining techniques to the catalogue of photometric sources observed by the Sloan Digital Sky Survey (SDSS) [1]. The development of this approach based on the extraction of knowledge from massive datasets, to this specific astronomical problem has yielded two main results: a) to provide cutting-edge scientific results; b) to demonstrate that effective research can be conducted in astronomy using data mining by a practical example of the application clustering and pattern recognition algorithms. The whole scientific workflow has been developed and implemented inside the DAME (DAta Mining and Exploration) collaboration. DAME is a Service Oriented Architecture based technological platform that will provide a general-purpose and cross-disciplinary computationally distributed environment for knowledge extraction and data mining. It will offer to the astronomical community a large spectrum of statistical tools and computational facilities to produce science by exploiting the wealth of available massive astronomical datasets and powerful statistical algorithms.

◇ Astronomical Observatory of Trieste - INAF, laurino@oats.inaf.it
†Harvard-Smithsonian Center for Astrophysics, dabrusco@head.cfa.harvard.edu
‡Department of Physical Sciences - University of Naples, longo@na.infn.it
♮ Astronomical Observatory of Capodimonte (Naples) - INAF, brescia@na.astro.it.

2. The overall picture: candidate astronomy and data mining

Classical astronomy has heavily relied on spectroscopic observations for several tasks, such as classifying sources and determining their redshift. Even though spectroscopy today is yet fundamental to gain insight into the physical processes which determine the luminous emission of astronomical sources, the unprecedented abundance of accurate photometric observations for very large samples of sources has spurred the development of the so called "candidate astronomy", i.e. a branch of astronomy which exploits mostly photometry to accomplish tasks which usually require spectroscopic data. The loss of accuracy and effectiveness in the classification and redshift determination based on photometric information only, for example, is balanced in the candidate astronomy by a very accurate evaluation of the uncertainties affecting the estimates. On the other hand, the principles of candidate astronomy have become widely accepted when dealing with specific problems, as the determination of the spatial distribution of visible matter on very large scale. In such cases, the statistical tools used to characterize the description of the distribution of matter are specifically designed to minimize the effects of the uncertainties on the parameters derived solely from photometry and take advantage of the significantly larger samples of sources that candidate astronomy provides with respect to astronomy based on spectroscopic data only.

The availability of both photometric and spectroscopic observations for limited samples of sources observed in the modern all-sky astronomical surveys, has allowed the development of techniques which combine the information that spectroscopy provide to the application of such knowledge to the overwhelming majority of photometric sources. The only hypothesis that has to be verified, in these cases, is that the sample of sources with spectroscopic observations, usually called base of knowledge (BoK), is statistically representative of the bulk of sources with only photometric observations, i.e. that the BoK samples homogeneously the underlying population of photometric sources.

Quasars are the most energetic sources in the Universe, but still today we fall short of obtaining a complete census of their spatial distribution because of the their low apparent luminosity, as they sit at the farthest outposts of the observed Universe. The classical and most reliable method for the selection of quasars entails spectroscopic observations in the optical bands, which are available in limited number because they are time consuming. Much larger samples of quasars can be obtained by selecting quasars directly from photometric data (in this sense, candidate quasars) by exploiting the existence of a "well behaved" sample of quasars for which we have spectra and have been classified using spectroscopy (see, for example, the approach described in [9]). The effectiveness of this approach, which can be ranked as one of the most successful applications of the techniques of the candidate astronomy, can be evaluated, from an astronomical standpoint, using the BoK. The ability of the method to select only likely quasars and few wrong sources is expressed by the efficiency $e = N_{cand}^{conf}/N_{cand}$, with N_{cand} as the number of candidates quasars extracted from a given sample and N_{cand}^{conf} is the number of spectroscopically confirmed quasars among the candidates selected; the performance of the method in selecting the largest fraction possible of the confirmed quasars which are known to be in a given sample of photometric sources is measured, on the other hand, by the completeness $c = N_{cand}^{conf}/N_{conf}^{total}$, where N_{conf}^{total} is the total number of spectroscopically confirmed quasars in the sample considered and N_{cand}^{conf} is the number of confirmed quasars effectively selected as candidate quasars. The ideal extraction would yield unitary efficiency and completeness.

Our goal was the design of a comprehensive method for the extraction of candidate quasars from the catalogue of stellar sources in the SDSS survey and the determination of their three-dimensional positions in the redshift space based on data mining techniques and two different spectroscopic BoKs. Our approach is based on three different steps, namely the characterization of the distribution of point-like sources with spectroscopic classification from SDSS catalogue in the optical photometric parameter space; the extraction of photometric candidate quasars based on such characterization and the de-projection of the positions of the candidate quasars using photometric redshifts.

On the methodological side, this project is the prototypical challenge that data mining methods can successfully tackle in the field of candidate astronomy, since it involves different steps all related to the extraction of information available inside the distribution of large sample of sources in multidimensional parameter spaces, like the determination of clusters and the reconstruction of patterns. From the point of view of data mining, a very schematic representation of the process is the following:

- Unsupervised clustering of the first BoK (all spectroscopic stellar sources in the SDSS catalogue) in the photometric parameter space: this steps leads to the selection of an optimal clustering in terms of the overall separation between confirmed quasars and stars contained in the clusters;
- Geometrical characterization of the clusters containing mostly quasars and extraction of new candidate quasars based on such characterization;
- Supervised clustering of the second BoK (spectroscopically confirmed quasars with available redshift estimates) in the optical photometric parameter space : this step is necessary to improve the accuracy of the final photometric redshift reconstruction;
- Pattern recognition inside each distinct clusters for the determination of the photometric redshifts;

Our approach to these tasks is based on algorithms specifically designed, one of which (the Weak Gated Experts) has not been published yet. In any case, none of the data mining methods used have never been applied to astronomical problems. In the following section we will describe in detail the steps of the process outlined above and discuss the results.

3. Extraction of candidate quasars

The first part of our approach to the reconstruction of the three-dimensional distribution of candidate quasars in the redshift space consists in the characterization of the BoK distribution in the photometric parameter space and the extraction of candidate quasars from photometric data on the basis of such characterization (for details see [6]). The characterization of the BoK in the photometric parameter space is carried out by performing unsupervised clustering to the BoK distribution of sources in the parameter space. The spectroscopic information available for the members of the BoK is then used to label the clusters produced in the process and to perform the extraction of the photometric candidate quasars. The method can be summarized as follows. The distribution of sources belonging to the first BoK (i.e. the complete set of SDSS stellar sources with spectroscopic observations and classification) in the photometric parameter space is partitioned in separate groups of nearby sources by determining a specific clustering in the photometric parameter space. Such clustering is optimal in the sense that it maximizes the total separation between spectroscopically confirmed quasars and other sources (mainly stars) contained in a subset of clusters where quasars largely outnumber not-quasars. In general, a nominal "purity" threshold is defined (in the experiment described here an 80% threshold was applied). For each single clustering (i.e. different set of the clusters) produced by our procedure, each cluster is assigned to one of the three categories: "goal-successful", "notgoal-successful" and "unsuccessful" clusters. The clusters whose ratio between the number of confirmed quasars and the total number of member exceeds the threshold are called goal-successful, while, in the opposite case when the ratio between the number of not-quasars and the total number of members is higher than the threshold, the clusters are called notgoal-successful clusters and when neither these conditions are verified, clusters are assigned to the unsuccessful class. While the composition of the goal-successful and notgoal-successful clusters is fixed, all sources associated to the unsuccessful clusters are used as BoK for a second generation of clustering which determines a second generation of goal-successful and notgoal-successful clusters. The clustering process is repeated on unsuccessful clusters members until the percentage of sources associated to such class of clusters is lower than 5% of the number of sources in the original BoK. In this sense, the whole clustering procedure is recursive and is designed to extract all the information

contained in the distribution of the spectroscopic sample of sources. The optimal clustering is the clustering for which the overall separation between quasars and not-quasars is the largest and the fraction of quasars and not-quasars respectively in the goal-successful and notgoal-successful clusters is the highest (i.e., such clusters are "purest"); this separation can be expressed quantitatively by a parameter called the Normalized Success Ratio (NSR):

$$NSR = \frac{\sum_{i=1}^{N} \frac{n_i^{goal}(quasars)}{n_i^{goal}(total)} + \sum_{j=1}^{M} \frac{n_j^{notgoal}(notquasars)}{n_j^{notgoal}(total)}}{(N+M)\left(\frac{\sum_{k=1}^{N+M} n_k(total)}{\sum_{i=1}^{N} n_i^{goal}(quasars) + \sum_{j=1}^{M} n_j^{notgoal}(notquasars)}\right)} \tag{1}$$

where M and N are the numbers of goal-successful and notgoal-successful clusters respectively, $n_i^{goal}(quasars)$ and $n_i^{goal}(total)$ are the numbers of quasars and total members of the i-th goal-successful clusters while $n_j^{notgoal}(notquasars)$ and $n_j^{notgoal}(total)$ are the numbers of not-quasars and total members of the j-th notgoal-successful cluster. This value is defined to be larger than one and tends to unity for clustering more and more similar to the "ideal" clustering, i.e. the situation where each cluster is perfectly pure and contains only members of the corresponding class. A simpler diagnostic of the performance of the selection of goal-successful clusters only, called total efficiency e_{tot}, is defined as:

$$e_{tot} = \sum_{i=1}^{N} \frac{n_i^{goal}(quasars)}{n_i^{goal}(total)} = \sum_{i=1}^{N} e_i \tag{2}$$

where e_i is the efficiency in the i-th goal-successful cluster. We have determined empirically that the final efficiency of the clustering process is not influenced by the purity threshold, while effecting the number of generations of clustering needed to reduce the fraction of sources associated to the unsuccessful clusters. Once the optimal clustering has been determined, the goal-successful clusters are used for the extraction of new candidate quasars. Every source belonging to the photometric dataset is projected in the photometric parameter space and associated to the closest cluster, be it goal or notgoal. If such cluster is goal-successful, the photometric source is selected as candidate quasar. The distances between the photometric point and the clusters are measured according to the Mahalanobis' distance [8]. By definition, the Mahalanobis' distance of multivariate vector $\vec{x} = (x_1, x_2, x_3, ..., x_N)$ from a distribution of points with mean located in the position $\vec{\mu} = (\mu_1, \mu_2, \mu_3, ..., \mu_N)$ and covariance matrix S is:

$$D_M(\vec{x}) = \sqrt{(x-\mu)^T S^{-1} (x-\mu)} \tag{3}$$

In the case the covariance matrix of the distribution of point is unitary, the Mahalanobis' distance reduces to the euclidean distance, while for a diagonal covariance matrix it is the so called normalized euclidean distance, defined for two random vectors $\vec{x} = (x_1, x_2, x_3, ..., x_N)$ and $\vec{y} = (y_1, y_2, y_3, ..., y_N)$, as:

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{\sigma_i^2}} \tag{4}$$

where σ_i is the standard deviation of the x_i over the sample set. From a geometrical point of view, the Mahalanobis' distance has the quality of taking into account the anisotropy of the distribution of members of the clusters.
The workflow describing the process of determination of the clusters and selection of the candidate quasars can be described as follows:

- The initial distribution of sources (first BoK) in the observed photometric parameter space is used as input for the first clustering algorithm;
- The first algorithm projects the BoK from the original parameter space to the latent space, where separation of clusters is maximally efficient;
- The first algorithms performs clustering in the latent space;
- The pre-clusters produced by the first algorithm in the latent space are projected back to the observed parameter space and sent to the second algorithm;
- The second algorithm produces a dendrogram (i.e. tree structure) of clusters (figure 3.1);
- The optimal clustering (in the sense explained above) of the dendrogram is selected;
- Mahalanobis' distance is used to extract candidate quasars by associating each photometric source to the closest clusters;

The workflow depicted here is general: even though we used two specific algorithms to perform the candidate quasars extraction, in principle any combination of a supervised and an unsupervided clustering algorithms can be used. We refer to *unsupervised* clustering as a clustering for which the number of clusters is not known *a priori*. For projection in the latent space and supervised clustering we used the Probabilistic Principal Surfaces (a method for dimensionality reduction which can be used also to perform clustering), and the Negative Entropy Clustering for the following agglomerative hierarchical unsupervised clustering. Both these different methods do not require any *a priori* hypothesis regarding the nature of the underlying distribution of sources, except for the initial number of pre-clusters produced by the PPS, but we have empirically proved that the final clustering, selecting according the procedure described above, do not change as a function of this parameter (if the number of pre-clusters given is "large" relative to final number of clusters produced in the optimal clustering).

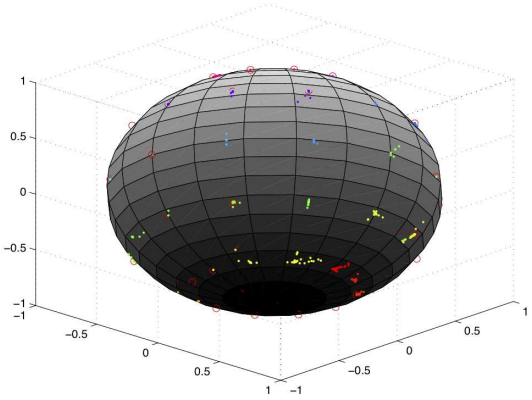
3.1. **The Probabilistic Principal Surfaces algorithm.** The Probabilistic Principal Surfaces model [5] belongs to the family of the so called latent variables methods and can be regarded as an extension of the Generative Topographic Mapping [2]. The goal of any latent variable model is to express the distribution $p(\mathbf{t})$ of the variable $\mathbf{t} = (t_1, \ldots, t_D) \in \mathbb{R}^D$ in terms of a smaller number of latent variables $\mathbf{x} = (x_1, \ldots, x_Q) \in \mathbb{R}^Q$ where $Q < D$. In order to achieve it, the joint distribution $p(\mathbf{t}, \mathbf{x})$ is decomposed into the product of the marginal distribution $p(\mathbf{x})$ of the latent variables and the conditional distribution $p(\mathbf{t}|\mathbf{x})$ of the data variables given the latent variables. It is convenient to express the conditional distribution as a factorization over the data variables, so that the joint distribution becomes:

$$p(\mathbf{t}, \mathbf{x}) = p(\mathbf{x})p(\mathbf{t}|\mathbf{x}) = p(\mathbf{x}) \prod_{d=1}^{D} p(t_d|\mathbf{x}) \tag{5}$$

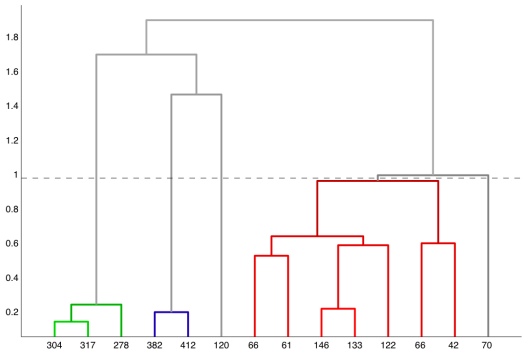
The conditional distribution $p(\mathbf{t}|\mathbf{x})$ is then expressed in terms of a mapping from latent variables to data variables, so that:

$$\mathbf{t} = \mathbf{y}(\mathbf{x}; \mathbf{w}) + \mathbf{u} \tag{6}$$

where $\mathbf{y}(\mathbf{x}; \mathbf{w})$ is a function of the latent variable $\mathbf{x}$ with parameters $\mathbf{w}$, and $\mathbf{u}$ is an $\mathbf{x}$-independent noise process. If the components of $\mathbf{u}$ are uncorrelated, the conditional distribution for $\mathbf{t}$ will factorize as in (5). From the geometrical point of view, the function $\mathbf{y}(\mathbf{x}; \mathbf{w})$ defines a manifold in the data space given by the image of the latent space. The definition of the latent variable model needs to be completed by specifying the distribution $p(\mathbf{u})$, the mapping $\mathbf{y}(\mathbf{x}; \mathbf{w})$, and the marginal distribution $p(\mathbf{x})$. The type of mapping $\mathbf{y}(\mathbf{x}; \mathbf{w})$ determines the specific latent variable model. The desired model for the distribution $p(\mathbf{t})$ of the data is then obtained by marginalizing over the latent variables:



(a) Schematic representation of the spherical manifold in the three dimensional latent space R^3 with the projection of the points distribution onto different nodes on the spherical surface.



(b) An example of dendrogram used as a representation of an agglomerative clustering process performed by the NEC algorithm.

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{x})p(\mathbf{x})d\mathbf{x}. \tag{7}$$

PPS define a non-linear, parametric mapping $\mathbf{y}(\mathbf{x};\mathbf{W})$, where $\mathbf{y}$ is defined continuous and differentiable, which projects every point in the latent space into a point in the data space. Since the latent space is Q-dimensional, these points will be confined to a Q-dimensional manifold non-linearly embedded into the D-dimensional data space. This implies that data points projecting near a principal surface node (i.e., a Gaussian center of the mixture) have higher influences on that node than points projecting far away from it. In order to estimate the parameters $\mathbf{W}$ and β we used the Expectation–Maximization (EM) algorithm [7]. In a $3D$ latent space, then, a spherical manifold can be constructed using a PPS with nodes $\{\mathbf{x}_m\}_{m=1}^{M}$ arranged regularly on the surface of a sphere in $\mathbb{R}^3$ latent space, with the latent basis functions evenly distributed on the sphere at a lower density. The motivation behind such a spherical manifold is that spherical PPS are particularly well suited to capture the sparsity and periphery of data in large input spaces [2]. After a spherical PPS model is fitted to the data, the data themselves are projected into the latent space as points onto a sphere (figure (1(a))).

Spherical PPS can be used as a "reference manifold" for classifying high-dimensional data. A reference spherical manifold is computed for each class during the training phase. In the test phase, a datum previously unseen by the network is classified to the class of its nearest spherical manifold. Obviously, the concept of "nearest" implies a distance computation between a data point $\mathbf{t}$ and the nodes of the manifold. Before this computation, the data point $\mathbf{t}$ must be linearly projected onto the manifold. Since a spherical manifold consists of square and triangular patches, each one defined by three or four manifold nodes, what is computed is an approximation of the distance. The Nearest Neighbor approximation method has been used because it allows to evaluate distances of each data point in the feature space to all nodes embedded in the spherical manifold. We want to emphasize that the non-linear relation between the features and the latent variables evaluated by PPS cannot be expressed mathematically in a closed form since it is completely empirical in nature. A representation of this relation can be recovered observing the positions of the same groups of objects in both the original parameter space and in the latent space, after the projection onto the 2-dimensional surface embedded in the latent space.

3.2. **The Negative Entropy Clustering algorithm.** Most unsupervised methods require the number of clusters to be provided *a priori*. This circumstance represents a serious problem when exploring large complex data sets where the number of clusters can be very high or, in any case, largely unpredictable. A classical agglomerative clustering algorithm is completely specified by assigning a definition of distance between clusters and a linkage strategy, i.e. a rule according to which clusters separated up to a given value of the distance are merged and others are not. Independently from the choice of the distance definition, successive generations of merging are carried out using updated distances between clusters: the resulting structure of clusters can be represented by the dendrogram (figure 3.1) until some convergence or threshold criterion is satisfied. A strictly geometrical interpretation of distances between clusters in the parameter space can be relaxed in order to generalize this class of algorithms, so that the distance between clusters can be defined as a generic function of the composition of the clusters in terms of parameter space coordinates. The NEC model is a hierarchical clustering algorithm based on the so called Negative Entropy, which measures the resemblance of a distribution of points to a multivariate Gaussian. This method uses Fisher's linear discriminant which is a classification method that first projects high-dimensional data onto a line, and then performs a classification in the projected one-dimensional space [2]. The differential entropy H of a random vector $\vec{y} = (y_1, ..., y_n)^T$ with a density f is defined as:

$$H(\vec{y}) = \int f(\vec{y}) \log f(\vec{y}) d\vec{y} \tag{8}$$

so that negentropy J can be defined as:

$$J(\vec{y}) = J(\vec{y}_{Gauss}) - H(\vec{y}) \tag{9}$$

where $\vec{y}_{Gauss}$ is a Gaussian random vector of the same covariance matrix as y. The Negentropy can be interpreted as a measure of non-Gaussianity and, since it is invariant for invertible linear transformations, finding an invertible transformation that minimizes the mutual information is roughly equivalent at finding directions in which the Negentropy is maximized. The only *a priori* information needed by NEC is a particular scalar value of the Negentropy called dissimilarity threshold T. We suppose to have n D-dimensional pre-clusters X_i with $i = 1, \ldots, n$ that have been determined by the PPS; these clusters are passed to NEC which ascertains whether each couple of contiguous clusters (according to the Fisher's linear discriminant) can or cannot be more efficiently modeled by one single multivariate gaussian distribution. This method can be easily generalized to other model distributions; we preferred to use a multivariate gaussian model only because the normal distribution can be considered a good approximation of any reasonably shaped peaked distribution. The optimal set of clusters (in the sense discussed in paragraph 3) is obtained for a given value of Negentropy, called the critical value of the dissimilarity threshold T_{cr}, by cutting the dendrogram in figure 3.1 with an horizontal line at $T = T_{cr}$.

4. Photometric redshifts determination

The techniques for the photometric redshift estimation rely upon the fact that the spectrum of radiation being emitted by most objects have strong features that can be detected by broad band filters. Even though the accuracy of the photometric redshift reconstruction, in general, is worse than the accuracy achieved with spectroscopic data, photometric redshifts provide a convenient way to estimate redshifts for large samples of sources for which expensive spectroscopy is not available. The physical mechanism responsible of the correlation between the photometric features and the redshift of an astronomical source mechanism implies a non-linear mapping between the photometric parameter space of the galaxies and the redshift values. Such highly non linear relation can be reconstructed, among other methods, using data mining. The family of "empirical" methods for the determination of photometric redshifts can be applied when a suitable BoK composed by sources

with accurate photometric observations is available, supplemented by spectroscopic redshifts for a subsample statistically representative of the parent population of photometric sources.
Neural networks (NN) may be very effectively used to reconstruct the relation between the "parameters" and the "target". In this specific case, the parameters are photometric measures of the extragalactic sources while the targets, an independent and reliable estimate of the quantity the NN are trained to evaluate, are redshifts of the sources measured from their observed spectra. The performance of the reconstruction of photometric redshift can be evaluated by measuring the scatter of the distribution of the difference variable $\Delta z = z_{phot} - z_{spec}$, i.e. the difference between the photometric and spectroscopic redshifts of all sources belonging to test set extracted from the BoK.

4.1. **The Multi Layer Perceptron.** Feed-forward neural networks provide a general framework for representing non linear functional mappings between a set of input variables and a set of output variables. One can achieve this goal by representing the non linear function of many variables by a composition of non-linear activation functions of one variable. A multi-layer perceptron may be represented by a graph: the input layer is made of a number of perceptrons equal to the number of input variables; the output layer, on the other hand, will have as many neurons as the output variables. The network may have an arbitrary number of hidden layers which in turn may have an arbitrary number of perceptrons. In a fully connected feed-forward network each node of a layer is connected to all the nodes in the adjacent layers. Each connection is represented by an adaptive weight which represents the strength of the synaptic connection between neurons. Generally speaking, along with the regular units, a feed-forward network presents a bias parameter for each layer. The bias parameter of the k-th layer is added to the activation function input of all the nodes in the $k+1$-th layer.
We will consider a generic feed-forward network with d input units, c output units and M hidden units in a single hidden layer. We shall refer to this kind of network as a two-layer network, counting the number of connection layers instead of the number of perceptron layers. The output of the j-th hidden unit of the k-th layer is formed by firstly computing the weighted sum of the inputs:

$$a_j^{(k)} = \sum_{i=0}^{d} w_{ji}^{(k)} z_i^{(k-1)} \tag{10}$$

where $w_{ji}^{(k)}$ denotes the weight related to the connection from the $k-1$-th layer to the j-th node of the k-th layer, and z_i is the activation state of the unit. Notice that the sum runs from 0 to d, and that we have included the bias parameter in the $k-1$-th units as $w_{j0}^{(k-1)} = b_{k-1}$, with a constant activation state $z_0^{(k-1)} = 1$. Then, the output of the j-th unit of the k-th layer is:

$$z_j^{(k)} = g\left(a_j^{(k)}\right) \tag{11}$$

where $g()$ is the activation function. By combining all these functions through the network, we obtain the outputs. For the k-th output unit, and assuming that the output activation function is linear ($\tilde{g}(a) = a$), we have:

$$y_k = \sum_{j=0}^{M} w_{kj}^{(2)} g\left(\sum_{i=0}^{d} w_{ji}^{(1)} x_i\right). \tag{12}$$

We shall refer to the topology of an MLP and to the weights matrix of its connections as to the model. In order to find the model that best fits the data, one has to provide the network with a set of examples: the training phase thus requires a BoK, also referred to, in this case, the training set.

4.2. **Weak Gated Experts.** Photometric redshifts estimation can be considered a regression problem. Regression is the task of predicting the dependent variable $d \in \mathbb{R}^N$ from the input vector $\mathbf{x} \in \mathbb{R}^M$ consisting of M random variables taken from an unknown and noisy input distribution. In many real world problems there is not a single mapping function from the parameter space to the target space of spectroscopic redshifts, so that a single MLP, or an MLP committee, cannot be sufficient for an accurate reconstruction of the color-redshift relation. Although a single global model can in principle approximate any function, even if it is piecewise defined, in real world problems the extraction of such global model from the data in the presence of degeneracies can very hard. Also, there is not a single global noise regime throughout the parameter space: the noise regime changes in different regions of the photometric parameter space. The input and target noise depend on the magnitudes of the sources so, in turn, depending on their distance from the observer, which is the information encoded in the redshift itself. Since the distribution of colors of the sources depend on the distance, it can be safely assumed that the noise depends on the input as well. Finally, the sparseness of the BoK of the quasars itself depends on the distribution of sources in the color space. Trying to learn the mapping function on different input space regions with different noise levels and different densities by a single network is a mismatch since the network could extract features that do not generalize well in some regions (local over-fitting), whilst it does not learn all it could potentially learn in the other regions (local under-fitting). In other terms, since the cost function is unique for a single network, a local over-fitting in some region may be compensated by a local under-fitting in other regions. For this reasons we implement a more complex architecture, following the mixture of experts paradigm. The basic idea of experts is to try and learn several local models from the data, in different regions of the parameter space. These experts are specialized over their subdomain and their outputs are linearly combined to form the output. The first proposed models that implement this idea are referred to as mixtures of experts. Gated experts are different from normal expert, since they non-linearly combine non-linear experts. The input space is also nonlinearly split into subspaces. The gating network is trained to learn both the segmentation of the input space and the input dependent coefficients $g_i(\mathbf{x})$ that combine the system outputs $y_i(\mathbf{x})$:

$$y = \sum_{i=1}^{K} g_i(\mathbf{x}) y_i(\mathbf{x}) \tag{13}$$

Each expert i is a standard neural network that learns a function $y_i(\mathbf{x})$ by means of a sigmoidal activation function hidden layer and a linear activation function output layer. The gating network, instead, has a classification flavor, with the K nodes of the output layer having a softmax activation function:

$$g_j = \frac{e^{s_i}}{\sum_{i=1}^{K} e^{s_i}} \tag{14}$$

where $s_i(\mathbf{x})$ is the output of the nodes of the hidden layer. The outputs of the gating networks are normalized to unity and their values express the competition among different experts, which is soft, in the sense that each input pattern has a non zero probability of being in the domain of each expert. This problem cannot be addressed by a supervised learning method only since, in general, we don't have any *a priori* knowledge on how to carve the input space. Thus, a complex cost function has to be derived to take into account all the variables. The cost function may be analytically derived [10]:

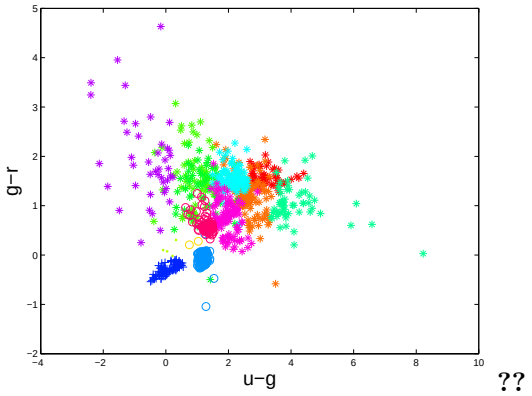
- The cost function cannot be minimized with gradient descent, but the problem can be reformulated and addressed by means of an Expectation Maximization algorithm;
- In order to find a consistent solution we need to assume that one and only one expert is responsible for each pattern. In other terms, we need to be sure that there is a way of isolating different sub-processes throughout the parameter space. For the quasars sample considered, this assumption is false.

We have elaborated a new method that can be viewed as a weak version of the gated experts since our method, while trying to take advantage of the gated experts strengths, also takes into account our knowledge of the problem from a physical point of view. Also, it is computationally cheaper than an Expectation Maximization method. In order to partition the input space, a fuzzy version of a simple clustering algorithm, namely the fuzzy k-means, or c-means was used. The classical crispy k-means algorithm finds, given the number of clusters k and a metric, the centroids that minimize the distance with the objects belonging to their clusters and maximize the distance among them, by an iterative method. When convergence is reached, each point in the input space belongs to one and only one cluster. c-means works exactly like its crispy counterpart for finding cluster centroids, except that in this case each source has a membership probability different from zero for every cluster. This property allows us to build clusters with soft boundaries, introducing some redundancy in the datasets, because we allow that some of the same patterns belong to different clusters. The gated experts need to be combined by means of a non-linear superposition: this task should be performed by a Maximization Expectation procedure together with the carving of the input space. Our weak gating network emulates the standard gating network by means of an MLP network in a regression configuration: this time, the features are the colors and the outputs of the experts. This means that, according to the colors of the sources, the gating network will try and learn how to recognize the patterns in the redshift estimations of the individual experts and assign to each source the best estimate of the photometric redshift.

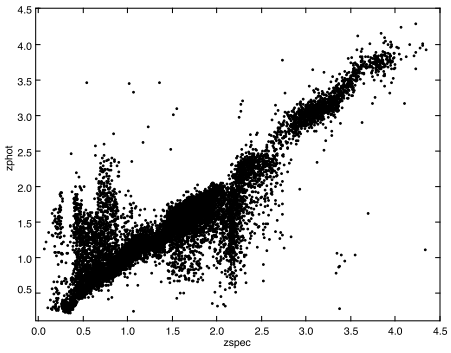
5. Results

The final result of the procedure described above is the largest catalogue of photometric candidate quasars to this date, extracted from the SDSS database, with photometric redshifts calculated with the WGE for every candidate. The ability of this method to discriminate between clusters in the multi-dimensional parameter space that would not be recognized by looking at the single color-color plane classically used to extract photometric candidate quasars is evident in figure **??**, where the final set of goal-successful clusters produced during the experiment from which the candidate quasars from the optical SDSS database were extracted, are plotted in different colors and symbols. While few groups os sources appear to be isolated even in this low dimensional projection, in the densest part of the plot the presence of distinct clusters is hardly discernible. A not up-to-date version of the catalogue of photometric candidate quasars without photometric redshift information is currently available online from the DAME collaboration website (for more details see [6]). The extraction procedure of candidate quasars from SDSS optical parameter space yielded a total efficiency $e_t \approx 95\%$ and a total completeness $c_t \approx 93\%$ for sources brighter than $m = 20.7$ in the SDSS r band. The efficiency and completeness were evaluated as the NSR value derived by the BoK and then cross-validated by applying the method to a control sample of sources not used previously for the clustering but with spectroscopic classification as well.

In a forthcoming version of the catalogue, the photometric redshifts and the associated error will be accompanied by quality flags which take into account the reliability of the estimation of the photometric redshifts as assessed through the same WGE workflow. The scatter plot of the spectroscopic redshifts against the photometric redshifts of the sources belonging to the test set extracted from the BoK used to train the WGE algorithm is shown in figure 1(d). The accuracy of the photometric redshifts is evaluated by calculating the median absolute (MAD) as robust measure of the spreading of the difference variable $\Delta z = z_{phot} - z_{spec}$. The MAD values together with the percentages of photometric redshifts differing less than 0.1, 0.2 and 0.3 respectively from the spectroscopic redshifts, are reported for our method and for other results in the astronomical literature in table 1. The best experiment performed (and used to produce the catalogue of photometric redshifts of the candidate quasars) leads to a $\sigma_{MAD} = 0.14$, which is similar if not better than the best results currently available in the astronomical literature, independently of the method employed. Since the WGE method is general and can be applied to different classes of extragalactic sources, the table 1 summarizes the results of the photometric redshift reconstruction in few different cases involving



(c) Distribution of the BoK in the color-color plane generated by u-g and g-r colors. Distinct goal-successful clusters are plotted with different symbols and colors.



(d) Scatter plot of the spectroscopic and photometric redshifts calculated with the WGE algorithm for a subset of the BoK used for the training phase.

TABLE 1. MAD estimates and fraction of sources with difference variables smaller than 0.1, 0.2 and 0.3 for four different photometric redshifts reconstruction experiments. In the dataset column, **S** stands for optical quasars candidates extracted from the SDSS, **G** stands for GALEX colors. The last three columns are the fraction of sources for which the photometric redshifts differ from the spectroscopic redshifts less then 0.1, 0.2 and 0.3 (0.01, 0.02 and 0.03 for galaxies).

Dataset	Sources	σ_{MAD}	$\Delta_{0.1(0.01)}$	$\Delta_{0.2(0.02)}$	$\Delta_{0.3(0.03)}$
S	quasars	0.14	48.7%	70.3%	78.9%
SG	quasars	0.09	67.9%	85.4%	91.0%
SG	quasars	0.10	62.9%	83.6%	90.0%
S	galaxies	0.02	40.6%	68.6%	83.9%

both candidate quasars and galaxies in a 5-dimensional optical parameter space obtained by SDSS photometry and adding photometric information from the GALEX ultraviolet bands [4].

6. DAME

The crucial role played by the multi-disciplinary expertise needed to deal with the ongoing burst of data complexity and to perform data mining and exploration on Massive Data Sets (MDS) has been recently certified by the constitution, within the IVOA (International VO Alliance), of an interest group on Knowledge Discovery in Data Bases (KDD-IG) which is seen as the main interface between the IVOA technical infrastructure and the VO enabled science. In this context, the DAME project intends: (a) to provide the VO with an extensible, integrated environment for Data Mining and Exploration; (b) to support the VO standards and formats, especially for application interoperability; (c) to abstract the application deployment and execution, so to provide the VO with an opaque general purpose computing platform taking advantage of the modern technologies (e.g. Grid, Cloud, etc...). Data Mining can be considered as the frontier of VObs enabled science since it represents the only way to capture and reveal the scientific knowledge (patterns, trends, correlations, etc.) hidden behind the complexity of MDS. The DAME project aims at creating a distributed e-infrastructure to guarantee integrated and asynchronous access to algorithms apt to mine data collected by very different experiments and scientific communities in order to correlate them and improve their scientific usability. The project consists of a data mining framework with powerful software instruments capable of working on MDS in a distributed computing environment.

So far, the VObs effort has focused on the realization of the low-level tools and on the definition of standards. Our project extends this fundamental target by integrating it in an infrastructure, joining service-oriented software and resource-oriented hardware paradigms, including the implementation of advanced tools for KDD purposes. Furthermore, the DAME design takes into account the fact that the average scientists cannot and/or does not want to become an expert also in Computer Science or in the fields of algorithms and Information Technology. In most cases the r.m.s. scientist (our end user) already owns his own algorithms for data processing and analysis and has implemented private routines/pipelines to solve specific problems. The KDD scheme adopted in the DAME package is based on Soft Computing methods, belonging to the typical dichotomy of machine learning methods which confronts supervised and unsupervised methods. All these methods have a common data mining paradigm: the AI (Artificial Intelligence) technique as self-adaptive exploration methodology. DAME is currently available as a working prototype at http://dame.na.infn.it, where users can perform tasks like the photometric redshifts estimation as described above. It has been successfully used also for educational purposes, allowing data mining unaware researchers to acquaint themselves with supervised data mining applications to astronomy and astrophysics within the Virtual Observatory framework. This prototype is just a proof of concept of what DAME is meant to be, which is currently in an advanced stage of development. Much effort has been devoted, during the design phase, to enforce flexibility and extendibility. The core DAME Service Oriented Architecture can be extended in many directions: for example, one can write his own client, or can provide a new driver for a different deployment environment (for both computing and storage). Also, one can develop low level Data Mining classes to enrich the low level Data Mining Models, or he can develop high level plugins implementing specific science cases which make use of the underlying Data Mining Models.

References

[1] K. N. Abazajian and e. a. Adelman-McCarthy. The Seventh Data Release of the Sloan Digital Sky Survey. *ApJS*, 182:543–558, June 2009.

[2] C. M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, Oxford, UK, 1996.

[3] K. Borne. Astroinformatics: A 21st Century Approach to Astronomy Research and Education. In *Bulletin of the American Astronomical Society*, volume 41 of *Bulletin of the American Astronomical Society*, pages 578–+, Jan. 2010.

[4] T. Budavári, S. Heinis, A. S. Szalay, M. Nieto-Santisteban, J. Gupchup, B. Shiao, M. Smith, R. Chang, G. Kauffmann, P. Morrissey, D. Schiminovich, B. Milliard, T. K. Wyder, D. C. Martin, T. A. Barlow, M. Seibert, K. Forster, L. Bianchi, J. Donas, P. G. Friedman, T. M. Heckman, Y. Lee, B. F. Madore, S. G. Neff, R. M. Rich, and B. Y. Welsh. GALEX-SDSS Catalogs for Statistical Studies. *ApJ*, 694:1281–1292, Apr. 2009.

[5] K.-Y. Chang and J. Ghosh. A unified model for probabilistic principal surfaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(1):22–41, 2001.

[6] R. D'Abrusco, G. Longo, and N. A. Walton. Quasar candidates selection in the Virtual Observatory era. *MNRAS*, 396:223–262, June 2009.

[7] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[8] P. Mahalanobis. On tests and measures of group divergence. Theoretical formulae. *J. Proc. Asiat. Soc. Bengal*, pages 541–588, 1936.

[9] G. T. Richards, A. D. Myers, A. G. Gray, R. N. Riegel, R. C. Nichol, R. J. Brunner, A. S. Szalay, D. P. Schneider, and S. F. Anderson. Efficient Photometric Selection of Quasars from the Sloan Digital Sky Survey. II. ˜1,000,000 Quasars from Data Release 6. *ApJS*, 180:67–83, Jan. 2009.

[10] A. S. Weigend, M. Mangeas, and A. N. Srivastava. Nonlinear gated experts for time series: Discovering regimes and avoiding overfitting, 1995.