Various the completeness and the efficiency, also

Science, Astronomy

Various data mining algorithms have been applied by astronomers in like most of the different applications in astronomy. But long-term researches and several mining projectshave been made by experts in this field of data mining making use of data related to the study of astronomy because astronomy has created numerous magnificent datasets that are flexible to the approach along withnumerous other areas like as medicine and high energy physics. Instances of suchnumerous projects are the SKICAT-Sky Image Cataloging and Analysis System for catalogproduction and analysis of the catalog from digitized sky surveys particularly the scans given by the second Palomar ObservatorySky Survey; the JAR Tool- Jet Propulsion Laboratory Adaptive Recognition Tool used for recognition of volcanoes formed in over 30, 000 images of Venus which came by the Magellanmission; the following and more general Diamond and the Lawrence Livermore National Laboratory Sapphire project work. Object classification Classification is an crucial preliminary step in the scientific method as it provides a way for arranging information in a method that may be used to make hypotheses and compare easily with models. The two most useful concepts in objectclassification are the completeness and the efficiency, also known as recall and precision.

They are generally defined in terms of true and false positives(TP and FP) and true and false negatives (TN and FN). The completeness is the fraction of those objectsthat are in reality of a given type that are classified as that type: and the efficiency is the fraction of objects generally classified as a given typethat are truly of that type These two quantities are interesting astrophysically because, while one wants both higher completeness and efficiency, there is mostly a tradeoff involved. The importance of each often mostly depends on the application, for instance, an investigation of such rare objects generallyrequires high completeness while allowing some contamination (lower efficiency) but statistical clustering ofcosmological objects requires high efficiency even at the cost of completeness. Star-Galaxy Separation Due to their physical size in comparison to their distance from us, almost all the stars are unresolved in photometric datasets, and therefore appear as pointsources. Galaxies despite being furtheraway, generally subtend a larger angle and appear as extended sources. However, other astrophysicalobjects such as quasars and supernovae, are also seen as as point sources.

Thus, the separation of photometric catalog into starsand galaxies, or more generally, stars, galaxies and otherobjects, is an importantproblem. The number of galaxies and stars in typical surveys (of order 108 or above) requires that such separation must beautomated. This problem is a well studied one and automatedapproaches were employed before current data mining algorithms became famous, for instance, during digitization done by the scanning of variousphotographic plates by machines such as the APM and DPOSS. Severaldata mining algorithms have been applied, including ANN, DT, mixturemodelling and SOM with most algorithms achieving over efficiency around 95%.

Typically, this is performed using a set of measured morphological parametersthat are made from the survey photometry, with perhaps colors or other information, such as the seeing. Theadvantage of data mining approach is that all such information abouteach object is easily incorporated. Galaxy Morphology Galaxies come in a rangeof numerous sizes and shapes, or more collectively, morphology. The most well-known system for the morphological classification of galaxies is the Hubble Sequence of elliptical, spiral, barredspiral, and irregular, along with various subclasses. This system correlates to many physical properties known to be crucial in the formation and formation of galaxies. Because galaxy morphologyis a tough and complex phenomenon that correlates to the underlying the subject of physics, but is notunique to any one given process, the Hubble sequence has shown, despiteit being rather subjective and based on visible-light morphology originally created from blue-biased photographic plates. The Hubble sequence has been extended in various othermethods, and for data miningpurposes the T system has been extensively taken into consideration.

This system maps the categorical Hubble types E, S0, Sa, Sb, Sc, Sd, and Irr onto the numerical values -5 to 10. One can train a supervised algorithm to allot T types to images for which measured parameters are made available. Such parameters can be completely morphological, or comprise of other information such as color. Aseries of paperswritten by Lahav and collaborators doexactly the same, by applying ANNs to predict the T type of galaxies at low redshift, and finding equal amount of accuracy tohuman experts. ANNs have also been applied to higher redshift data to distinguish betweennormal and unique galaxies and the fundamentally topologicaland unsupervised SOM ANN has been used to classify various galaxies from Hubble Space Telescope images, where the initial distribution of variousclasses is unknown. Likewise, ANNs have been used to obtain the morphological types from galaxy spectra. Photometric redshifts Anarea of astrophysics that has greatly increased in popularity in the last few years is the estimation of redshifts from photometric data (photo-zs). This is because, although the distances are less accurate than the ones obtained withspectra, the sheer numberof objects with photometric measurements can often make up for the reduction in individualaccuracy by suppressing thestatistical noise of an ensemble calculation.

The two common approaches to photo-zs are the template method andthe empirical training the set method. The template approach has manydifficult issues, including calibration, zero-points, priors, multi-wavelength performance(e. g., poor in the mid-infrared), and difficulty handlingmissing or incomplete training data. We focus in this review on theempirical approach, as it is an implementation of supervised learning. 3. 2. 1.

Galaxies At low redshifts, the calculation of photometric redshifts for normal galaxies is quite straightforward due to the break in the typical galaxy spectrum at 4000A. Thus, as a galaxy is redshifted withincreasing distance, the color (measured as a difference in magnitudes) changesrelatively smoothly. As a result, both template and empiricalphoto-z approaches obtain similar outcomes, aroot-mean-square deviation of ~ 0.

02 in redshift, which is near to the best possible result giventhe intrinsic spread in the properties. This has beenshown with ANNs SVM DT, kNN, empirical polynomial relations, numerous template-based studies, and several other procedures. Athigher redshifts, acheiving accurate results becomes more difficult because the 4000A break is shifted redward of the optical, galaxies are fainter and thus spectral data are sparser, and galaxies intrinsically evolve over time. While supervised learning has been successfully used, beyond the spectral regime the obvious limitation arises that in order to reach thelimiting magnitude of the photometric portions of surveys, extrapolation wouldbe required. In this regime, or where only small training sets are available, template-based results can be used, but without spectral information, thetemplates themselves are being extrapolated.

However, the extrapolation of thetemplates is being done in a more physically motivatedmanner. It is likely that the more general hybrid method of using empirical data to iteratively improve the templates or the semi-supervised proceduredescribed in will ultimately provide a more elegant solution. Anotherissue at higher redshift is that the available numbers of objects can becomequite small (in the hundreds or fewer), thus reintroducing the curse of dimensionality by a simple lack of objects in comparison to measured wavebands. The methods of dimension reduction can help to mitigate this effectVarious data mining algorithms have been applied by astronomers in like most of the different applications in astronomy. But long-term researches and several mining projectshave been made by experts in this field of data mining making use of data related to the study of astronomy because astronomy has created numerous magnificent datasets that are flexible to the approach along withnumerous other areas like as medicine and high energy physics. Instances of suchnumerous projects are the SKICAT- Sky Image Cataloging and Analysis System for catalogproduction and analysis of the catalog from digitized sky surveys particularly the scans given by the second Palomar ObservatorySky Survey; the JAR Tool- Jet Propulsion Laboratory Adaptive Recognition Tool used for recognition of volcanoes formed in over 30, 000 images of Venus which came by the Magellanmission; the following and more general Diamond and the Lawrence Livermore National Laboratory Sapphire project work. Object classification Classification is an crucial preliminary step in the scientific method as it provides a way for arranging information in a method that may be used to make hypotheses and compare easily with models.

The two most useful concepts in objectclassification are the completeness and the efficiency, also known as recall and precision. They are generally defined in terms of true and false positives(TP and FP) and true and false negatives (TN and FN). The completeness is the fraction of those objectsthat are in reality of a given type that are classified as that type: and the efficiency is the fraction of objects generally classified as a given typethat are truly of that type These two quantities are interesting astrophysically because, while one wants both higher completeness and efficiency, there is mostly a tradeoff involved. The importance of each often mostly depends on the application, for instance, an investigation of such rare objects generallyrequires high completeness while allowing some contamination (lower efficiency) but statistical clustering ofcosmological objects requires high efficiency even at the cost of completeness. Star-Galaxy Separation Due to their physical size in comparison to their distance from us, almost all the stars are unresolved in photometric datasets, and therefore appear as pointsources.

Galaxies despite being furtheraway, generally subtend a larger angle and appear as extended sources. However, other astrophysicalobjects such as quasars and supernovae, are also seen as as point sources. Thus, the separation of photometric catalog into starsand galaxies, or more generally, stars, galaxies and otherobjects, is an importantproblem. The number of galaxies and stars in typical surveys (of order 108 or above) requires that such separation must beautomated. This problem is a well studied one and automatedapproaches were employed before current data mining algorithms became famous, for instance, during digitization done by the scanning of variousphotographic plates by machines such as the APM and DPOSS. Severaldata mining algorithms have been applied, including

ANN, DT, mixturemodelling and SOM with most algorithms achieving over efficiency around 95%.

Typically, this is performed using a set of measured morphological parametersthat are made from the survey photometry, with perhaps colors or other information, such as the seeing. Theadvantage of data mining approach is that all such information abouteach object is easily incorporated. Galaxy Morphology Galaxies come in a rangeof numerous sizes and shapes, or more collectively, morphology. The most well-known system for the morphological classification of galaxies is the Hubble Sequence of elliptical, spiral, barredspiral, and irregular, along with various subclasses. This system correlates to many physical properties known to be crucial in the formation and formation of galaxies. Because galaxy morphologyis a tough and complex phenomenon that correlates to the underlying the subject of physics, but is notunique to any one given process, the Hubble sequence has shown, despiteit being rather subjective and based on visible-light morphology originally created from blue-biased photographic plates. The Hubble sequence has been extended in various othermethods, and for data miningpurposes the T system has been extensively taken into consideration. This system maps the categorical Hubble types E, S0, Sa, Sb, Sc, Sd, and Irr onto the numerical values -5 to 10.

One can train a supervised algorithm to allot T types to images for which measured parameters are made available. Such parameters can be completely morphological, or comprise of other information such as color. Aseries of paperswritten by Lahav and collaborators doexactly the same, by applying ANNs to predict the T type of galaxies at low redshift, and finding equal amount of accuracy tohuman experts. ANNs have also been applied to higher redshift data to distinguish betweennormal and unique galaxies and the fundamentally topologicaland unsupervised SOM ANN has been used to classify various galaxies from Hubble Space Telescope images, where the initial distribution of variousclasses is unknown. Likewise, ANNs have been used to obtain the morphological types from galaxy spectra.

Photometric redshifts Anarea of astrophysics that has greatly increased in popularity in the last few years is the estimation of redshifts from photometric data (photo-zs). This is because, although the distances are less accurate than the ones obtained withspectra, the sheer numberof objects with photometric measurements can often make up for the reduction in individualaccuracy by suppressing thestatistical noise of an ensemble calculation. The two common approaches to photo-zs are the template method andthe empirical training the set method. The template approach has manydifficult issues, including calibration, zero-points, priors, multi-wavelength performance(e. g., poor in the mid-infrared), and difficulty handlingmissing or incomplete training data.

We focus in this review on theempirical approach, as it is an implementation of supervised learning. 3. 2. 1. Galaxies At low redshifts, the calculation of photometric redshifts for normal galaxies is quite straightforward due to the break in the typical galaxy spectrum at 4000A. Thus, as a galaxy is redshifted withincreasing distance, the color (measured as a difference in magnitudes) changesrelatively smoothly. As a result, both template and empiricalphoto-z approaches obtain similar outcomes, aroot-mean-square deviation of ~ 0.

02 in redshift, which is near to the best possible result giventhe intrinsic spread in the properties. This has beenshown with ANNs SVM DT, kNN, empirical polynomial relations, numerous template-based studies, and several other procedures. Athigher redshifts, acheiving accurate results becomes more difficult because the 4000A break is shifted redward of the optical, galaxies are fainter and thus spectral data are sparser, and galaxies intrinsically evolve over time. While supervised learning has been successfully used, beyond the spectral regime the obvious limitation arises that in order to reach thelimiting magnitude of the photometric portions of surveys, extrapolation wouldbe required. In this regime, or where only small training sets are available, template-based results can be used, but without spectral information, thetemplates themselves are being extrapolated. However, the extrapolation of thetemplates is being done in a more physically motivatedmanner.

It is likely that the more general hybrid method of using empirical data to iteratively improve the templates or the semi-supervised proceduredescribed in will ultimately provide a more elegant solution. Anotherissue at higher redshift is that the available numbers of objects can becomequite small (in the hundreds or fewer), thus reintroducing the curse of dimensionality by a simple lack of objects in comparison to measured wavebands. The methods of dimension reduction can help to mitigate this effect