V. Kůrková et al. (Eds.): ITAT 2014 with selected papers from Znalosti 2014, CEUR Workshop Proceedings Vol. 1214, pp. 7–12
http://ceur-ws.org/Vol-1214, Series ISSN 1613-0073, © 2014 D. Bednárek, M. Kruliš, J. Yaghob, F. Zavoral

Employing Evolutionary Algorithms for Classification of Astrophysical Spectra

David Bednárek, Martin Kruliš, Jakub Yaghob, and Filip Zavoral

Parallel Architectures/Algorithms/Applications Research Group
Faculty of Mathematics and Physics, Charles University in Prague
Malostranské nám. 25, Prague, Czech Republic
{bednarek,krulis,yaghob,zavoral}@ksi.mff.cuni.cz

Abstract: In the past decade, automated astronomical observatories collected huge amounts of data which can no longer be explored by astronomers individually. In our case, we deal with optical spectra produced by multi-object low-resolution spectrographs. Due to lower resolution and higher level of noise in such surveys, individual spectra rarely offer reliable information; however, since many similar objects expectedly exist in the universe, global analysis of the spectrum database may reveal classes of objects sharing similar properties. In this paper, we propose a novel evolutionary approach to classification of spectral data which is expected to achieve finer level of detail than traditional methods. Furthermore, we describe the most computationally-intensive parts of the method in the form of parallel cache-aware algorithm.

Keywords: evolutionary algorithm, spectrum, classification, astrophysics, astroinformatics

1 Introduction

Studying the spectra of celestial objects was the key to many (if not the majority) of astronomical discoveries of the last two centuries and it still remains the most valuable instrument in stellar astronomy.

Stellar spectrum is a recording of radiation intensity in the frequency domain, usually over a range of visible or near-infrared wavelengths. The most prominent features of a stellar spectrum are its general shape (also called *continuum*), *absorption lines*, and *emission lines*. In most studies including our approach, the absorption lines are considered the most important features.

Spectra reveal significant clues about the chemical composition, temperature, and velocity of the observed object; however, the interpretation of the observed facts is difficult because different physical processes may result in similar observations.

Large surveys like the Sloan Digital Sky Survey (SDSS) [1] produce hundreds of thousands spectral measurements using multi-object spectrographs. These devices have lower resolution than single-object spectrographs by design; in addition, the nature of a large survey requires that relatively fainter objects are measured. Consequently, the measured spectra often lack enough detail required by traditional classification methods; in particular, individual measurement of absorption lines is possible only in the cases of most prominent lines.

This lack of detail may be balanced by a global approach where a model of the spectrum is compared to the measured intensity along the complete width of the spectrum, instead of focusing on the prominent lines only. Models of stellar spectra can either by synthesized from the theory or based on the measurement of well-studied prototype objects, including the Sun. Astronomers have created libraries of synthetic spectra with varying number of parameters [2, 3] . To match a real observation, the model parameters must be determined; a number of methods has been already proposed based on machine learning [4], principal component analysis [5], or combined methods [6].

Unfortunately, some scientifically interesting classes of objects like Be stars still lack sufficiently general models and their variability, given by their nontrivial geometry, makes parameter determination difficult even for simple models. Therefore, the classification of such objects into subclasses is still based rather on specific features observed in their spectra [7] than on the parameters of a matching physical model.

Our proposed approach is inspired by evolutionary algorithms. The goal of the method is to create synthetic spectra to match the observations; however, the synthetic spectra are not based on a physical model of the object. Instead, each synthetic spectrum is matched against as many observations as possible, trying to cover the given set of observations by as few synthetic spectra as possible. Of course, the observed objects are often similar but not completely identical. Therefore, each synthetic spectrum is allowed to undergo a transformation before matching to an observation; the evolutionary algorithm tries to evolve both the synthetic spectra and the transformation parameters at once.

Compared to the traditional meaning of the synthetic spectra, the physics in our model is greatly reduced: The set of the lines in our synthetic spectrum is not derived from the assumed chemistry of the object but simply placed to match the observations. On the other hand, the profiles of the lines are physically sound (corresponding to the effect of Doppler broadening), and also the transformations of the spectra correspond to physical variations like differences in temperature or radial velocity.

If such a synthetic spectrum is successfully matched to a set of observed objects, each of the matched observations require different parameters of the spectrum transformation. Consequently, the synthetic spectrum corresponds to a hypothetical object whose spectral characteristics are close to all the matched objects and the transformations required to match the individual objects are related to the difference between the hypothetical object and the observed object.

Our synthetic spectrum does not offer any physical model; however, if a model is assigned to the synthetic spectrum by other means, e.g., by the inspection by an astronomer, the model will probably apply to the observed objects as well. In addition, the physical meaning of the spectrum transformation allows the determination of the required change in parameters of the assigned model, including the verification whether the change is physically plausible. Nevertheless, the physical interpretation of the transformation is not a part of our method; the physical background of the transformation merely serves as a means of defining physically relevant notion of similarity.

The main purpose of the method is the reduction of the number of spectra that must be inspected manually; consequently, there is no strict requirement of separation of the resulting clusters, only the requirement for high intracluster similarity.

The paper is organized as follows. Section 2 revises related work in the fields of spectra classification and evolutionary algorithms. The following section describes the mathematical background of our synthetic spectra, their transformation, and how they are matched to the observations. Section 4 describes our co-evolution algorithm. Section 5 summarizes the approach and suggests the modes of its application.

2 Related Work

The idea of the astronomic spectra classification is not completely new. In the past, the most preferred approach was based on the examination of significant spectral line [8, 9].

Bazaghan [10] proposed the *self organizing maps* as an unsupervised artificial neural network algorithm for classification of the stellar spectra. Jiang et al. [11] used *principal component analysis* methods to reduce dimensionality of the data, where only the first two eigenvectors are selected. Furthermore, they proposed a hierarchical clustering method for the data mining approach.

Bromová et al. [7] attempted to employ wavelets as descriptors of the stellar spectra. The spectra were sampled by discrete wavelet transformation and various transformations of the coefficients into Euclidean space were used, thus the descriptors were simple vectors. The k-means algorithm [12] was applied on the descriptors to find similar spectra, especially to identify the spectra of Be stars. Their implementation achieved 76% precision on a sample set of 656 spectra with manually annotated ground truth; however, the sample set consisted of high-precision low-noise spectra from a single-object spectrograph. When applied to multi-object spectrograph measurements, the precision was lost.

A simpler technique [13] used 2D curves like the Bezier curve to approximate the histogram and then compare the coefficients or the defining points of the curves.

Evolutionary algorithms and especially genetic algorithms have been used for various types of classification and clustering problems. As a representative, we have selected the work of Maulik [14], which proposes a clustering technique based on genetic algorithm. The algorithm is similar to the k-means clustering algorithm [12], but the centroids are a population of individuals which is refined using the genetic approach.

In physics, genetic algorithms have been used for classification and pattern recognitions in mass spectra. Lavine et al. presented a genetic algorithm for classification of wood types measured by Raman spectroscopy [15]. A year later, Lavine presented more generalized version of the genetic algorithm for pattern recognition in mass spectra [16]. However, the aim of these methods is the classification into reliably defined and well separated classes of materials while our goals include discovering of such classes.

Another approach to mass spectra analysis was devised by Geurts et al. [17]. Their method is based on assembling a decision tree which detects proteomic biomarkers in the spectrum. Their objective was to devise a method for automated detection of various diseases in the body fluids.

3 Mathematic Model

Each synthetic spectrum S_i is defined as a set of lines; each line is determined by its position l, width w, and intensity d.

$$S_i = \{\langle l_{i,1}, w_{i,1}, d_{i,1}\rangle, ..., \langle l_{i,n_i}, w_{i,n_i}, d_{i,n_i}\rangle\}$$

The line parameters are expressed in units that allow easy application of physically relevant transformations: the position is expressed as the logarithm of wavelength because Doppler shifts act as multipliers of the wavelength, the width uses a unit corresponding to the temperature associated to Doppler broadening, and the intensity is measured using the logarithm of attenuation which is proportional to the density of the gas generating the absorption line.

Each line generates wavelength-dependent attenuation corresponding to Doppler broadening, described by the function

$$A_{l,w,d}(\lambda) = e^{-de^{-((log_{10}\lambda - l)/w)^2}}$$

Since computing the value of this function is expensive and cannot be vectorized, the function is tabulated in our implementation, using equidistant sampling in the four dimensions $log_{10}\lambda$, l, w, and d. In $log_{10}\lambda$, the sampling points are equal to the sampling points of the observed spectra (which are, fortunately, normalized to equidistant in our database). In l, w, and d, the approximation was improved by quadratic interpolation, i.e., using tabulated values of A and all its first and second partial derivatives. In addition, the following equivalences are used to reduce the number of samples:

$$A_{l+x,w,d}(\lambda) = A_{l,w,d}(\lambda \cdot 10^{-x})$$

$$A_{l,w,d_1+d_2}(\lambda) = A_{l,w,d_1}(\lambda) \cdot A_{l,w,d_2}(\lambda)$$

Using these tricks, the number of samples, required to achieve the precision similar to the precision of the measured spectra, was reduced to approx. $3 \cdot 10^6$. This amount could easily fit into the memory but still poses significant burden on the cache hierarchy, requiring careful design of the algorithm.

The transformation of the spectrum consists of a global change to all line parameters and multiplication of the resulting attenuation curve by a black-body radiation curve B_T for the temperature T. The resulting synthetic spectrum curve is defined as:

$$C_{i,T,l',w',d'}(\lambda) = B_T(\lambda) \cdot \prod_{j=1}^{n_i} A_{l_{i,j}+l',w_{i,j}\cdot w',d_{i,j}\cdot d'}(\lambda)$$

Thus, each transformed spectrum is defined by the quintuple $\langle i,T,l',w',d' \rangle$ where i is the index of base spectrum and $P = \langle T,l',w',d' \rangle$ are the parameters of the transformation

3.1 Matching Synthetic Spectra to Observations

Matching against the observed spectrum O incorporates another physically relevant transformation – multiplying the value by a factor m reflecting the both the absolute luminosity of the object and its distance from the observer. The value of m is determined using the method of least squares, i.e., minimizing the sum

$$R = \sum_{\lambda} \frac{(O(\lambda) - m \cdot C(\lambda))^2}{\sigma^2(\lambda)}$$

In this definition, the quadratic residua are divided by variance $\sigma^2(\lambda)$ estimated for each measured wavelength. The inverse variance $\frac{1}{\sigma^2(\lambda)}$ is a part of the SDSS data along the flux $O(\lambda)$. The application of estimated variance allows to suppress the parts of the measured spectra affected by the noise caused by the atmospheric background.

The minimal value of R is

$$R_{\mathtt{min}} = \sum_{\lambda} \frac{O(\lambda)^2}{\sigma^2(\lambda)} - \frac{(\sum_\lambda \frac{O(\lambda)\cdot C(\lambda)}{\sigma^2(\lambda)})^2}{\sum_\lambda \frac{C(\lambda)^2}{\sigma^2(\lambda)}}$$

Since we need a measure of the match quality which is consistent over differently luminous objects, we use a normalized form of the sum:

$$\Delta(O,C) = 1 - \frac{(\sum_\lambda \frac{O(\lambda)\cdot C(\lambda)}{\sigma^2(\lambda)})^2}{\sum_\lambda \frac{O(\lambda)^2}{\sigma^2(\lambda)} \cdot \sum_\lambda \frac{C(\lambda)^2}{\sigma^2(\lambda)}}$$

This function may act as a distance between the observed spectrum O and the synthetic spectrum S, albeit its symmetry is broken by the fact that the variance σ^2 is associated to the observation. Since the presence of the σ^2 factors is merely a technical trick to minimize the influence the sky background and it does not significantly affect the method, we will omit the σ^2 data in the description of the evolutionary algorithm, for simplicity.

4 Evolutionary Algorithm

Assume that we have a set of observations $O = \{O_1,...,O_m\}$.

As stated in the previous section, our goal is to establish a set of synthetic spectra $S = \{S_1,...,S_n\}$, and to assign one of the synthetic spectra to every observation together with a set of transformation parameters.

Nevertheless, the nature of evolutionary algorithms require a population of candidate solutions – in our case, it means that every observation may be assigned to several spectra from the set S, each with different transformation parameters.

Thus, our population consists of two parts: A set S of synthetic spectra and a set $P = \{P_1,...,P_p\}$ of *pairings*. Each pairing is a tuple

$$P_k = \langle i_k, j_k, T_k, l'_k, w'_k, d'_k \rangle$$

where i_k is the index of a base synthetic spectrum, j_k is the index of an observed spectrum, and $\langle T_k, l'_k, w'_k, d'_k \rangle$ are the parameters of the transformation as described in the previous section. In other words, P_k is a link between the synthetic spectrum S_{i_k} and the observation O_{j_k}.

The quality of each pairing P_k is evaluated using the previously defined distance function Δ:

$$q(P_k) = 1 - \Delta(O_{j_k}, C_{i_k,T_k,l'_k,w'_k,d'_k})$$

In traditional settings, the fitness function q would control the evolution of the population P and defining mutation and/of crossover operators over P members would be sufficient to create a working evolutionary algorithm. However, in our case, we must simultaneously evolve also the set S of the synthetic spectra.

The structure and flow of data is depicted in Figure 1.

4.1 Symbiotic Evolution

Our population consists of two parts, S and P, which shall evolve simultaneously like a pair of different species living in a common environment. In addition, the members

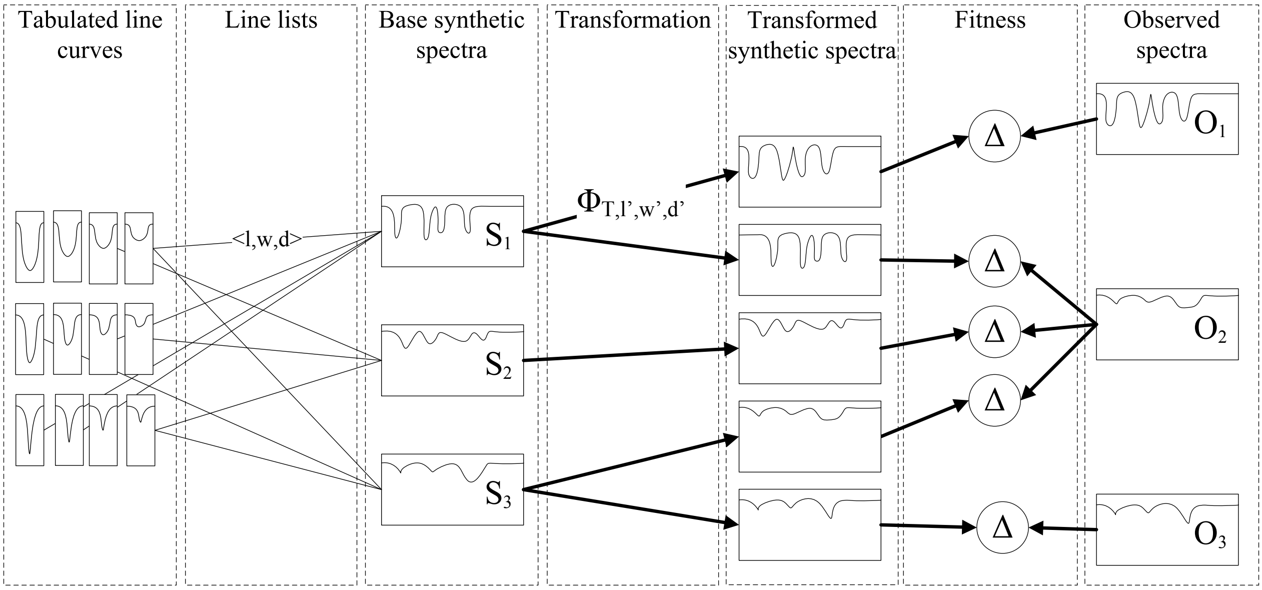

Figure 1: Synthetic spectra transformation and matching

of *P* are linked to members of *S*, resembling symbiosis between the two species. The symbiosis is asymmetric as each individual from *S* hosts several individuals from *P*.

Although there are numerous examples of such symbiosis in the nature, only few attempts [18] exist to transfer this mechanism to the world of evolutionary algorithms.

Such a symbiosis requires to solve a set of additional problems:

- If *S* was fixed, the population *P* could be divided into isolated islands attached to individual observations from *O*, evolving independently. However, the co-evolution of *S* causes that all members of *P* may mutually interfere through the *S* population.

- The transformation expressed by a *P* individual is not able to hit the associated observation perfectly; the minimal possible distance (however measured) from the observation depends on the associated *S* individual. Thus, some *P* individuals may approximate their objects more easily than others. Consequently, enforcing a global fitness function for *P* would prematurely kill individuals attached to those observations from *O* which are hard to approximate. In other words, there is no globally acceptable fitness function for *P* members; instead, *P* members must be compared only locally among the subset attached to the same observation.

- A fitness function must be defined for the members of *S*. Naturally, it shall be based on the fitness of the *P* members linked to the evaluated individual of *S*. However, merely summing the fitness will not work due to the expected large differences in the number of linked *P* members.

- Migration (i.e., re-linking a *P* member to a different *S* member) must be supported as an equivalent to mutation of the *P* member. Consequently, a notion of distance must be defined on *S* in order to favorize short-distance migrations over long-distance ones – i.e., small mutations over large ones.

- Poor-fitness *S* members must not die-out because *P* members may be linked to them.

In our approach, these problems are addressed as follows:

The system keeps track of family relationships in the set *S*. It means that, when an individual is created by mutation or crossover, the source individuals are preserved and the parent-child relation is saved in the form of a directed acyclic graph. Fortunately, the *S* individuals are shared and thus significantly less numerous than the *P* population; consequently, storing the complete history of its evolution is feasible.

Keeping *S* members forever solves the problem of orphaned *P* members. Nevertheless, the main advantage is elsewhere:

The graph of relationships allows distance measurement between the members of *S*, consistent with the factual difference of the corresponding synthetic spectra. If two members of *S* share a common ancestor, the number of generations between them and the ancestor may be used as a measure of distance. Assuming that the genetic operators represent movement to small distances in the space of spectra, close relatives in the *S* graph are close also in the space. Of course, the converse implication is not true, because distant nodes in the *S* graph may also represent neighbors in the spectral space.

This notion of distance is used in the mutation of *P* members: A *P* member may randomly relocate to a different S individual; however, only to a sibling or a child of its previous host – this way, the relocation preserves locality in the space of spectra.

In addition, the relocation of the P members is controlled by their fitness: Poorly fit individuals relocate to siblings in an attempt to find a replacement for the current poorly fitting host. Well fit individuals relocate to children of their host, attempting to improve the fitness.

The evolution of S members must support the need for relocation of P members. It means that S members occupied by a number of well-fit P members must generate offspring to enable their relocation. In other words, the fitness of an S member, being a controller of its fertility, must reflect the sole presence of well-fit P members, regardless of their number and independently of the presence of other P members.

This approach is detailed in Algorithm 1. The algorithm contains a number of steps that require tuning or may be realized in different ways. For instance, the vague final condition on line 3 may be implemented as a check for stagnating summary fitness over the P population; however, in many cases, it would be rather limited by the computing resources available.

Algorithm 1 The co-evolution algorithm

```
Require: $O$ the observations ; $N$, $t$, $M_{\mathtt{M}}$, $M_{\mathtt{X}}$ evolution parameters
 1: $S$ := random population of spectra
 2: $P$ := random population of pairings on $S \times O$
 3: while not satisfied do
 4:     compute the fitness $q(P_k)$ for every $P_k \in P$
 5:     for every $O_j \in O$, determine $P'_j$ = the top $N$ associated $P$ members according to $q$
 6:     $P' := \bigcup_j P'_j$
 7:     for every $S_i \in S$, determine $n(S_i)$ = the number of $P'$ members hosted by $S_i$
 8:     $S' := \{S_i \in S \mid n(S_i) > t\}$
 9:     for every $S_i \in S'$ generate $M_{\mathtt{M}}$ children of $S_i$ by random mutation
10:     randomly select $M_{\mathtt{X}}$ pairs from $S'$ for crossover
11:     relocate every $P_k \in P'$ to a randomly selected child of the linked $S$ member
12:     relocate every $P_k \in P \setminus P'$ to a randomly selected sibling of the linked $S$ member
13: end while
```

4.2 Parallel Implementation

The most computationally intensive part of the co-evolution algorithm is the calculation of population fitness. Thanks to the tabulation of line curves, the computation consists mostly of multiplication and addition. These simple operations are well supported by SIMD instructions of contemporary CPU's; consequently, the throughput of the arithmetic unit is very high, in the order of 10^{10} operations per second per core.

Given the high performance of the arithmetic unit, the memory and cache subsystem becomes the bottleneck of the algorithm. Furthermore, the observed spectra are matched against the base synthetic spectra almost randomly and the base spectra are also created from essentially randomly selected line curves. Consequently, iteration along the population would lead to random access both to the tabulated line curves and to the database of observed spectra. Thus, such a naive approach would lead to poor cache hit ratios and, consequently, poor performance.

To improve the performance of the fitness calculation, we developed the Algorithm 2. The algorithm is based on dividing the data into appropriately sized groups which can fit in a level of the memory hierarchy:

The data set O of observed spectra may be so large that it must be located in external storage. Consequently, it must be divided into groups $\{G^O_j\}$ and processed group-by-group. The size of every G^O_j group shall be selected so that the corresponding spectra fit into the last level of cache. For our Xeon CPUs with 8 MB L3 caches, the optimal group size was about 50 spectra.

The set of tabulated line curves is typically slightly larger than the last level of cache; therefore, it is divided into groups $\{G^A_n\}$. To manage the division, the lines that constitute the synthetic spectra must be collected and sorted according to the associated line curves (line 9 of Algorithm 2).

To employ parallelism while avoiding locking, every G^O_j group must be further divided into as many groups as there are computing units. To balance the size of the groups, the division is done indirectly, dividing the set of lines sorted along the observed spectra (line 12 and 13) into equivalently sized groups $\{G^L_m\}$.

5 Conclusion

Our evolutionary algorithm categorizes observations into sets represented by a common synthetic spectrum – this spectrum offers a reasonable representative of the set of observations.

Furthermore, the method may improve the comprehensibility of the spectrum, because the evolution of the synthetic spectrum produces results similar to the averaging of the observations in the associated set. Averaging measurements is a well-established technique used for improving the signal-to-noise ratio; however, raw averaging would produce invalid results due to differences in the measurement conditions like Doppler shifts. Our method helps to find the set to be averaged and, at the same time, it suggests transformations whose inversions shall be applied before averaging.

Although our approach is similar to clustering and similarity-based methods, there is a principal difference: Our method does not guarantee that similar objects will be arranged in the same set. There is only a complementary guarantee that the objects in the same set are similar.

Algorithm 2 Parallel fitness calculation algorithm

```
Require: $O$ the observations ; $A$ line curves ; $S$ synthetic spectra line lists ; $P$ pairings and transformation parameters
Ensure: fitness value $q(P_k)$ for every $P_k \in P$
for each group $G_j^O \subseteq O$ of observed spectra do
    read the group $G_j^O$ into memory
    $L' := \emptyset$
    for each pairing $P_k \in P$ associated to a spectrum from $G_j^O$ [in parallel] do
        allocate and initialize buffer $C_k$ for the transformed spectrum
        compute transformed line list $L_k'$ from the base line list and the transformation parameters
        $L' := L' \cup L_k'$
    end for
    sort $L'$ by the index of the referenced line curve
    for each group $G_n^A \subseteq A$ of the line curves do
        determine the range $L'' \subseteq L'$ corresponding to $G_n^A$
        sort $L''$ by the index of the observed spectrum
        for each group $G_m^L \subseteq L''$ [in parallel] do
            for each transformed line $\langle k, l, w, d \rangle \in G_m^L$ do
                multiply the line curve $A_{l,w,d}$ to the buffer $C_k$
            end for
        end for
    end for
    for each pairing $P_k \in P$ associated to a spectrum from $G_j^O$ [in parallel] do
        compute fitness $q(P_k)$ of $C_k$ w.r.t. the associated spectrum
    end for
end for
```

Even more, the sets may intersect, so the method must be perceived only as a means of reducing the number of observations to be inspected manually.

Acknowledgment

This paper was supported by Czech Science Foundation (GACR) projects P103/13/08195 and P103/14/14292P and by SVV-2014-260100.

References

[1] Sloan digital sky survey. [Online]. Available: http://www.sdss3.org/

[2] P. Coelho, B. Barbuy, J. Meléndez, R. Schiavon, and B. Castilho, "A library of high resolution synthetic stellar spectra from 300 nm to 1.8 μm with solar and α-enhanced composition," *Astronomy & Astrophysics*, vol. 443, pp. 735–746, 2005.

[3] A. Palacios, M. Gebran, E. Josselin, F. Martins, B. Plez, M. Belmas, and A. Lebre, "Pollux: a database of synthetic stellar spectra," *arXiv preprint arXiv:1003.4682*, 2010.

[4] O. Fuentes, "Automatic determination of stellar atmospheric parameters using neural networks and instance-based machine learning," *Experimental Astronomy*, vol. 12, no. 1, pp. 21–31, 2001.

[5] A. Recio-Blanco, A. Bijaoui, and P. De Laverny, "Automated derivation of stellar atmospheric parameters and chemical abundances: the matisse algorithm," *Monthly Notices of the Royal Astronomical Society*, vol. 370, no. 1, pp. 141–150, 2006.

[6] Y. Wu, A.-L. Luo, H.-N. Li, J.-R. Shi, P. Prugniel, Y.-C. Liang, Y.-H. Zhao, J.-N. Zhang, Z.-R. Bai, P. Wei *et al.*, "Automatic determination of stellar atmospheric parameters and construction of stellar spectral templates of the Guoshoujing telescope (LAMOST)," *Research in Astronomy and Astrophysics*, vol. 11, no. 8, p. 924, 2011.

[7] P. Bromová, P. Škoda, and J. Zendulka, "Wavelet based feature extraction for clustering of Be stars," in *Nostradamus 2013: Prediction, Modeling and Analysis of Complex Systems*. Springer, 2013, pp. 467–474.

[8] S. Veilleux and D. E. Osterbrock, "Spectral classification of emission-line galaxies," *The Astrophysical Journal Supplement Series*, vol. 63, pp. 295–310, 1987.

[9] J. Baldwin, M. Phillips, and R. Terlevich, "Classification parameters for the emission-line spectra of extragalactic objects," *Publications of the Astronomical Society of the Pacific*, pp. 5–19, 1981.

[10] M. Bazarghan, "Application of self-organizing map to stellar spectral classifications," *Astrophysics and Space Science*, vol. 337, no. 1, pp. 93–98, 2012.

[11] J. Bin, P. J. Chang, Y. Z. Ping, and G. Qiang, "A data mining application in stellar spectra," in *Computer Science and Computational Technology, 2008. ISCSCT '08. International Symposium on*, vol. 2, 2008, pp. 66–69.

[12] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Applied statistics*, pp. 100–108, 1979.

[13] F. Yamaguchi and F. Yamaguchi, *Curves and surfaces in computer aided geometric design*. Springer-Verlag Berlin, 1988.

[14] U. Maulik and S. Bandyopadhyay, "Genetic algorithm-based clustering technique," *Pattern recognition*, vol. 33, no. 9, pp. 1455–1465, 2000.

[15] B. K. Lavine, C. Davidson, A. J. Moores, and P. Griffiths, "Raman spectroscopy and genetic algorithms for the classification of wood types," *Applied Spectroscopy*, vol. 55, no. 8, pp. 960–966, 2001.

[16] B. K. Lavine, C. Davidson, and A. J. Moores, "Genetic algorithms for spectral pattern recognition," *Vibrational Spectroscopy*, vol. 28, no. 1, pp. 83–95, 2002.

[17] P. Geurts, M. Fillet, D. De Seny, M.-A. Meuwis, M. Malaise, M.-P. Merville, and L. Wehenkel, "Proteomic mass spectra classification using decision tree based ensemble methods," *Bioinformatics*, vol. 21, no. 14, pp. 3138–3145, 2005.

[18] A. Vahdat, M. I. Heywood, and A. N. Zincir-Heywood, "Symbiotic evolutionary subspace clustering," in *Evolutionary Computation (CEC), 2012 IEEE Congress on*. IEEE, 2012, pp. 1–8.