Astrophys Space Sci
DOI 10.1007/s10509-008-9816-5

ORIGINAL ARTICLE

Automated classification of sloan digital sky survey (SDSS) stellar spectra using artificial neural networks

Mahdi Bazarghan · Ranjan Gupta

Received: 14 November 2007 / Accepted: 21 April 2008

Abstract Automated techniques have been developed to automate the process of classification of objects or their analysis. The large datasets provided by upcoming spectroscopic surveys with dedicated telescopes urges scientists to use these automated techniques for analysis of such large datasets which are now available to the community. Sloan Digital Sky Survey (SDSS) is one of such surveys releasing massive datasets. We use Probabilistic Neural Network (PNN) for automatic classification of about 5000 SDSS spectra into 158 spectral type of a reference library ranging from O type to M type stars.

Keywords Probabilistic neural network · Spectra · Sloan digital sky survey · Spectral classification

1 Introduction

Intelligent systems based on pattern recognition tools like Artificial Neural Networks (ANNs) are now been used in various fields in applications like time series based event predictions, object classifications, data compression etc. In the past decade ANNs have found applications in classification of spectra, star-galaxy etc. which are typical Astronomical requirements where large data bases are now coming up. The Sloan Digital Sky Survey (York et al. 2000) is one of such publicly available source of data where the need of automatic classification of a large data set provides an ideal ground for ANNs. Previous studies involving spectral classification using ANNs include: Gulati et al. (1994a, 1994b, 1995), Von Hipple et al. (1994), Weaver and Torres-Dodgen (1995), Singh Harinder et al. (1998), Singh Harinder and Gupta (2003), Gupta et al. (2004), Bazarghan (2008) and also automated analysis of stellar spectra by Allende Prieto (2004). These pioneering efforts resulted in setting up the scene where ANNs have proven to be an established tool for classification of large set of stellar spectra.

We have used Probabilistic Neural Network Specht (1990) in MATLAB programming to classify 4999 test spectra of the SDSS into a set of 158 training spectra from a reference spectral library Jacoby et al. (1984). Probabilistic neural network is intrinsically a classifier with four layer network. Input layer is fully connected to the next layer that is the pattern layer. There is one pattern node for each training example in this layer. The next layer is a summation layer which sums the inputs from pattern units. The output layer with one neuron represents the maximum value in the summation layer. PNN uses a supervised training set to develop probability density functions within a pattern layer. As new pattern vectors are presented to the PNN for classification, they are serially propagated through the hidden layer by computing the dot product between the new pattern and each pattern stored in the hidden layer.

The result of classification of about 5000 SDSS stellar spectra is possibly a first attempt on this data set, though ANNs are already used on SDSS data for finding Galaxy types (Ball et al. 2004), measuring photometric redshifts (Vanzella et al. 2004) and separation of stars and galaxies (Qin et al. 2003).

Section 2 describes the SDSS spectral data set; Sects. 3 and 4 describe the PNN tool which has been applied to the

Electronic supplementary material The online version of this article (http://dx.doi.org/10.1007/s10509-008-9816-5) contains supplementary material, which is available to authorized users.

M. Bazarghan (✉) · R. Gupta
Inter University Center for Astronomy and Astrophysics,
Post Bag 4, Ganeshkhind, Pune 411007, India
e-mail: mahdi@iucaa.ernet.in

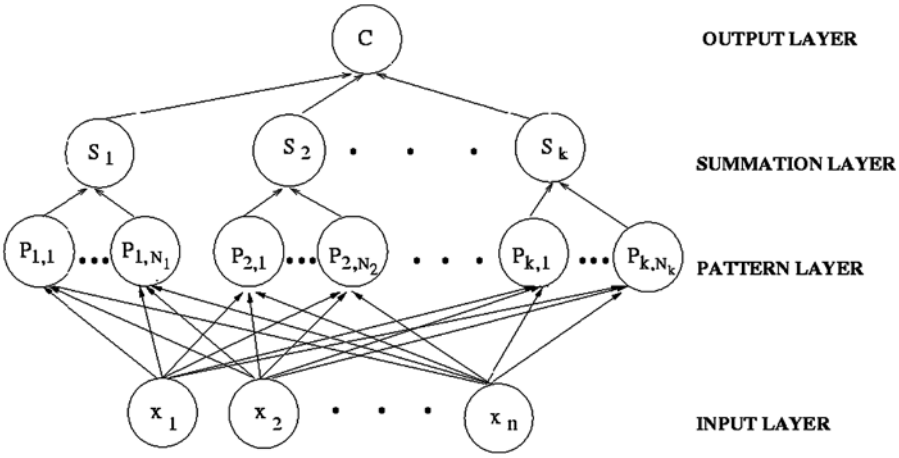

Fig. 1 Schematic of a typical probabilistic neural network

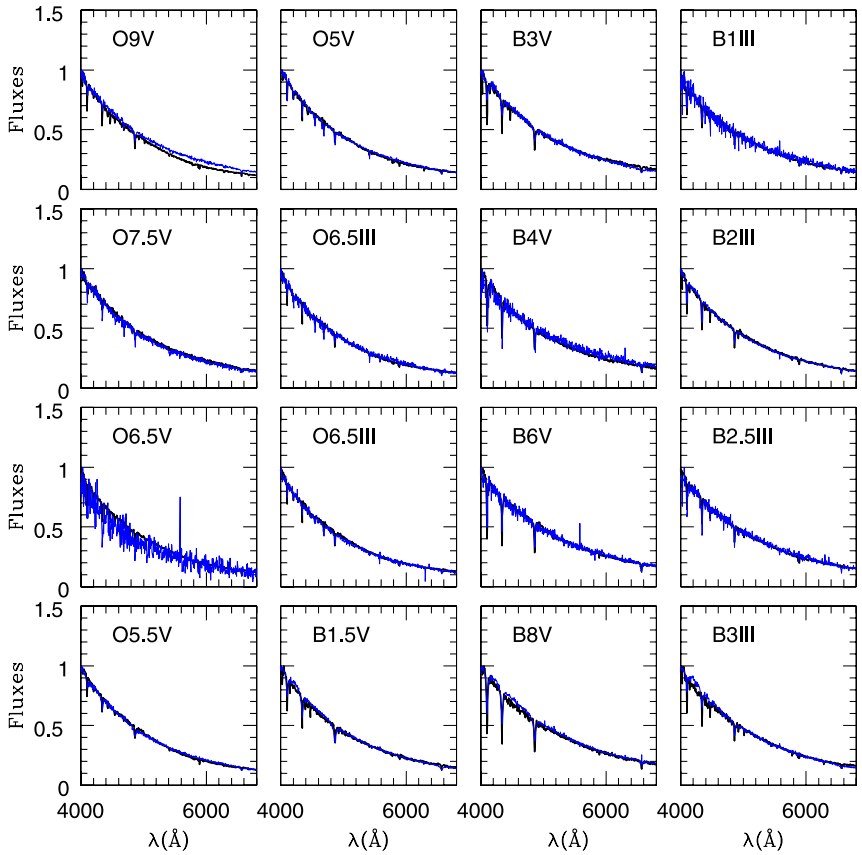

Fig. 2 SDSS-DR2 spectra with the corresponding JHC reference library spectra and its spectro-luminosity type (Table 1)

data set and the classification process details for this work and finally Sects. 5 and 6 give the results and discussion.

2 SDSS spectral data set and reference set of spectra

Sloan Digital Sky Survey is the most ambitious astronomical survey project ever undertaken. The survey will map in detail one-quarter of the entire sky with five broad band filters, determining the positions and absolute brightness of more than 100 million celestial objects. A technical summary of the survey is given in York et al. (2000). The data release to the community so far consists of the June 2001 Early Data Release (Stoughton 2002), the April 2003 Data Release One (Abazajian et al. 2003), the March 2004 Data Release Two (DR2) (Abazajian et al. 2004), the September 2004 Data Release Three (Abazajian et al. 2005), the June 2005 Data Release Four (Adelman-McCarthy et al. 2006), the June 2006 Data Release Five (Adelman-McCarthy et al. 2007a) and the June 2007 Data Release six (Adelman-McCarthy et al. 2007b).

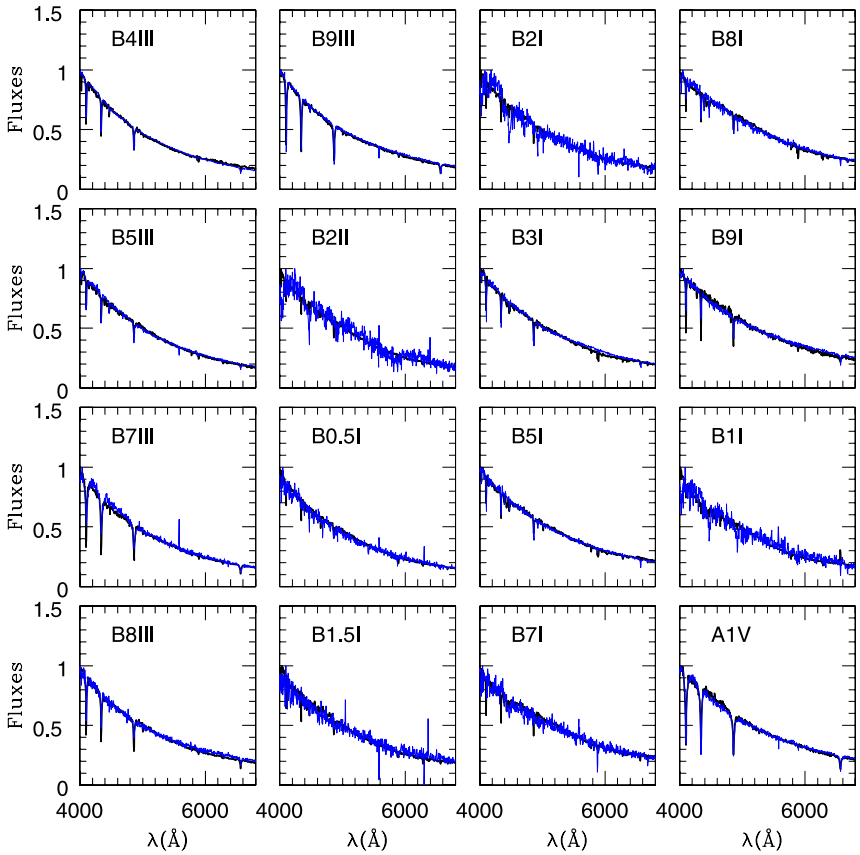

Fig. 3 SDSS-DR2 spectra with the corresponding JHC reference library spectra and its spectro-luminosity type (Table 1)

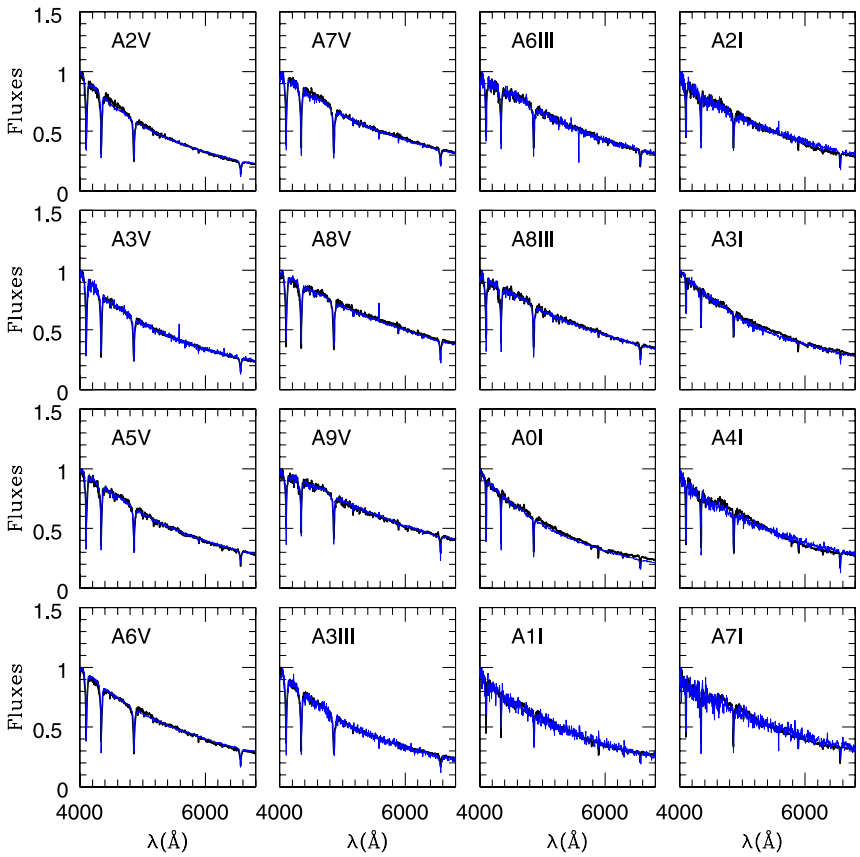

Fig. 4 SDSS-DR2 spectra with the corresponding JHC reference library spectra and its spectro-luminosity type (Table 1)

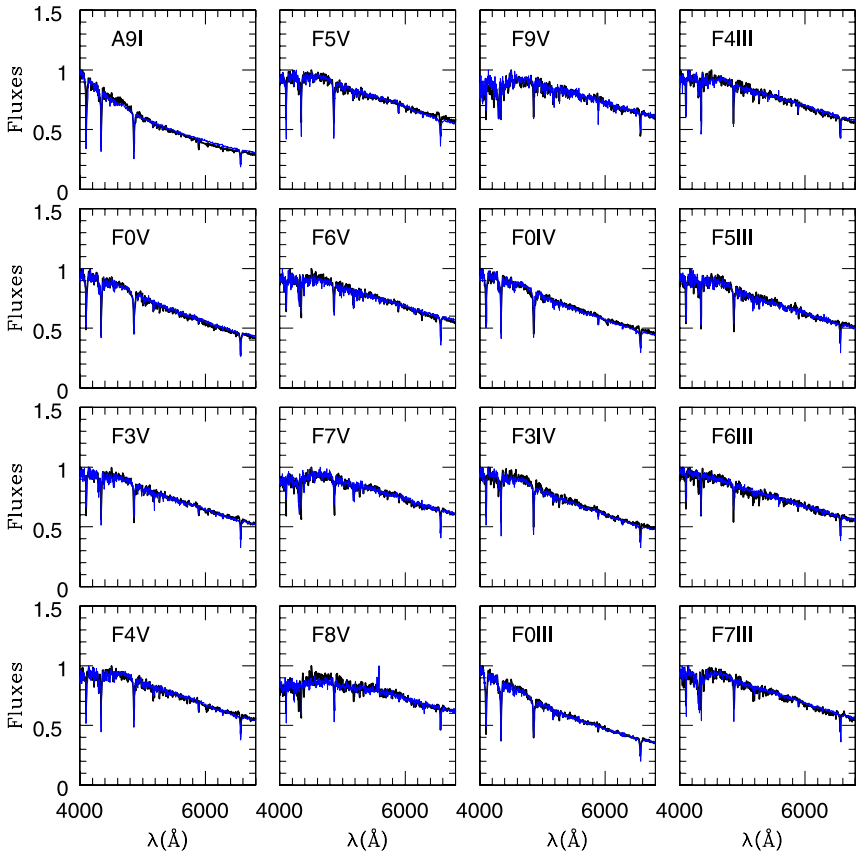

Fig. 5 SDSS-DR2 spectra with the corresponding JHC reference library spectra and its spectro-luminosity type (Table 1)

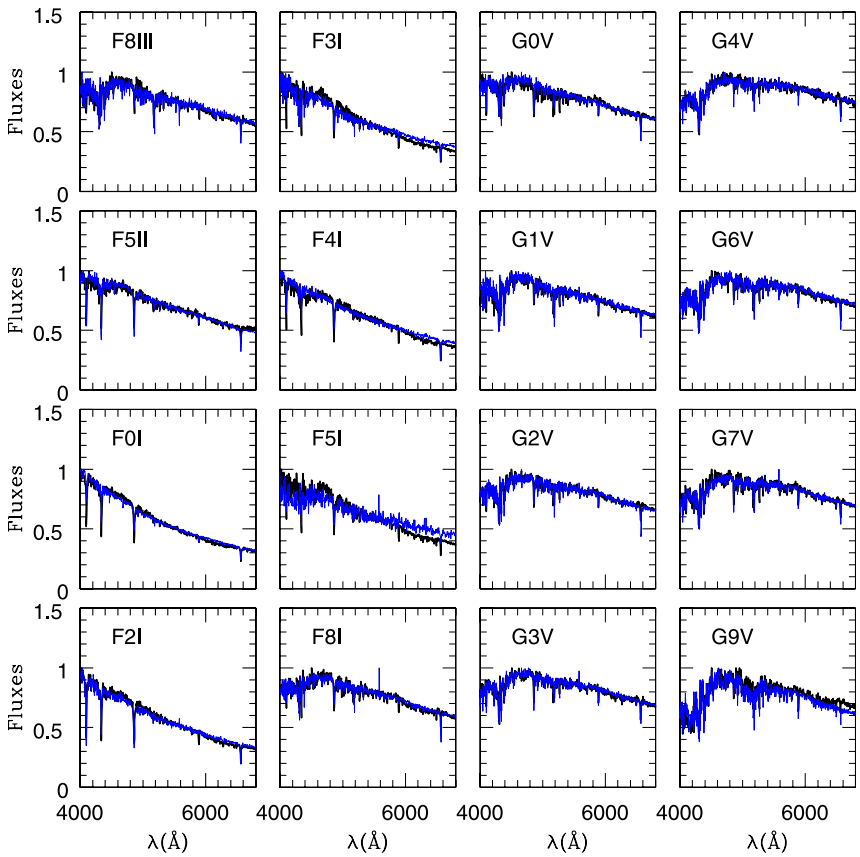

Fig. 6 SDSS-DR2 spectra with the corresponding JHC reference library spectra and its spectro-luminosity type (Table 2)

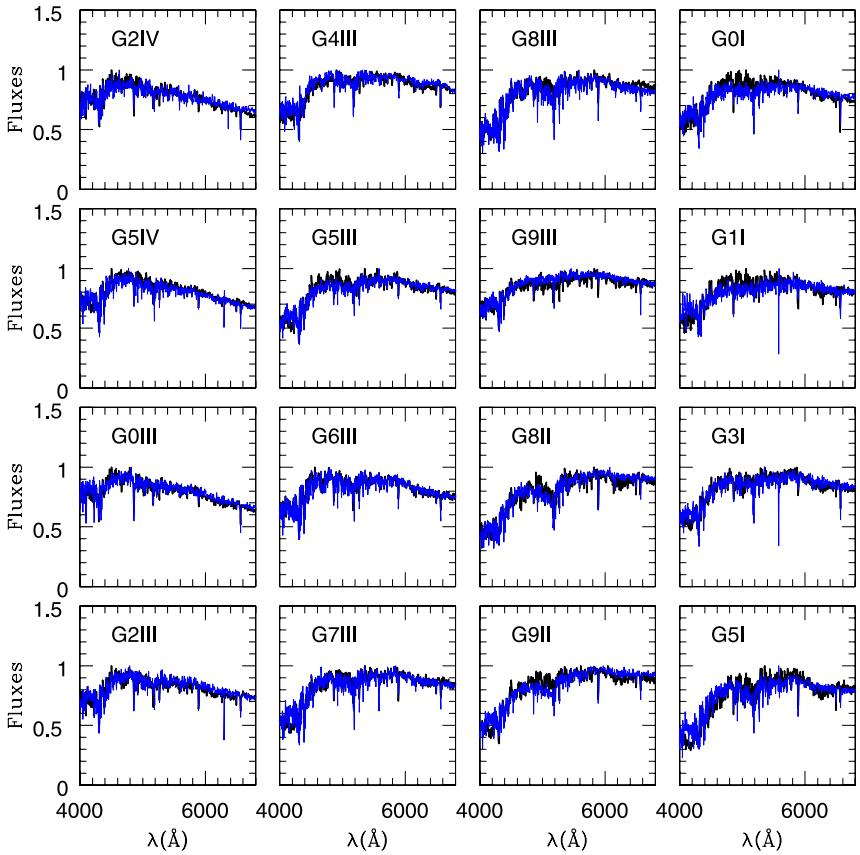

Fig. 7 SDSS-DR2 spectra with the corresponding JHC reference library spectra and its spectro-luminosity type (Table 2)

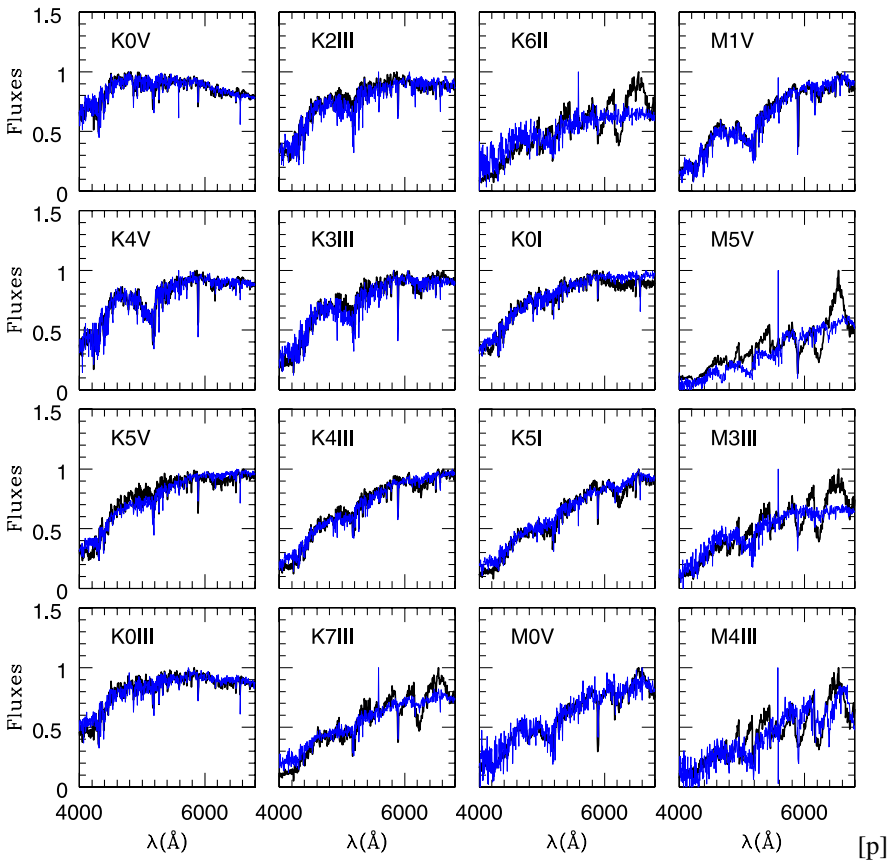

Fig. 8 SDSS-DR2 spectra with the corresponding JHC reference library spectra and its spectro-luminosity type (Table 2)

[p]

Table 1 List of spectra appearing in Figs. 2–5

SDSS spectra	Spectro-luminosity class	χ^2 value
spSpec-52367-0332-184	O9V	0.00142028
spSpec-51984-0279-228	O7.5V	0.00042054
spSpec-51986-0294-423	O6.5V	0.00701563
spSpec-51942-0301-626	O5.5V	0.00029600
spSpec-51673-0313-519	O5V	0.00029920
spSpec-52000-0335-478	O6.5III	0.00048330
spSpec-51633-0268-008	O6.5III	0.00042562
spSpec-51941-0272-510	B1.5V	0.00052162
spSpec-51909-0276-370	B3V	0.00062574
spSpec-51908-0277-521	B4V	0.00202127
spSpec-51689-0312-031	B6V	0.00089164
spSpec-51900-0278-242	B8V	0.00078749
spSpec-51615-0303-060	B1III	0.00133665
spSpec-51928-0291-105	B2III	0.00027330
spSpec-51665-0311-575	B2.5III	0.00082641
spSpec-51990-0340-352	B3III	0.00076806
spSpec-51658-0282-110	B4III	0.00022405
spSpec-51691-0350-366	B5III	0.00030136
spSpec-51658-0282-067	B7III	0.00111983
spSpec-51792-0354-196	B8III	0.00076665
spSpec-51691-0342-435	B9III	0.00045657
spSpec-51663-0315-624	B2II	0.00497184
spSpec-51818-0358-126	B0.5I	0.00156664
spSpec-52056-0329-201	B1.5I	0.00386845
spSpec-51688-0302-490	B2I	0.00440111
spSpec-51908-0277-055	B3I	0.00077680
spSpec-52370-0330-020	B5I	0.00050732
spSpec-51990-0310-393	B7I	0.00254126
spSpec-51671-0299-592	B8I	0.00088153
spSpec-51994-0293-108	B9I	0.00082505
spSpec-51789-0352-587	B1I	0.00405424
spSpec-51792-0354-451	A1V	0.00054832

Table 1 (*Continued*)

SDSS spectra	Spectro-luminosity class	χ^2 value
spSpec-51994-0309-313	A2V	0.00049425
spSpec-51789-0352-177	A3V	0.00062889
spSpec-51990-0310-307	A5V	0.00041014
spSpec-51658-0282-539	A6V	0.00040037
spSpec-51630-0266-233	A7V	0.00034089
spSpec-51994-0309-194	A8V	0.00058785
spSpec-51691-0350-044	A9V	0.00043968
spSpec-51994-0293-470	A3III	0.00072669
spSpec-51821-0359-263	A6III	0.00107148
spSpec-51821-0359-142	A8III	0.00043082
spSpec-51691-0342-518	A0I	0.00054997
spSpec-51662-0308-083	A1I	0.00183806
spSpec-51910-0269-055	A2I	0.00128973
spSpec-51883-0271-121	A3I	0.00061977
spSpec-51688-0302-167	A4I	0.00156048
spSpec-51662-0308-503	A7I	0.00258414
spSpec-51818-0358-078	A9I	0.00064258
spSpec-51959-0283-021	F0V	0.00046135
spSpec-51909-0276-426	F3V	0.00048153
spSpec-51662-0308-343	F4V	0.00047088
spSpec-51691-0342-357	F5V	0.00066779
spSpec-51818-0358-060	F6V	0.00057737
spSpec-52282-0328-287	F7V	0.00129746
spSpec-52375-0326-035	F8V	0.00240744
spSpec-52023-0287-188	F9V	0.00105487
spSpec-51703-0353-570	F0IV	0.00045788
spSpec-51699-0349-477	F3IV	0.00054928
spSpec-51692-0339-051	F0III	0.00034837
spSpec-51910-0275-458	F4III	0.00054288
spSpec-51908-0277-315	F5III	0.00099659
spSpec-51692-0339-214	F6III	0.00098363
spSpec-51612-0280-593	F7III	0.00145028

In order to have successful application of the ANN techniques and achieving effective results, one has to prepare the dataset well before applying to the neural network. They must be all uniform, having same wavelength scale, the starting and end wavelengths must be same for all the spectra and they must also have closely match spectral resolution. These must be valid to both the training and test dataset.

The training set for the PNN is a set of 158 spectra taken from Jacoby et al. (1984) (hereafter JHC which contains a total 161 spectra), which covers the wavelength range of 3510–7427 Å for various O to M type stars. The set of 4999 spectra from SDSS with wavelength range of 3800–9200 Å forms the test set which gets classified into 158

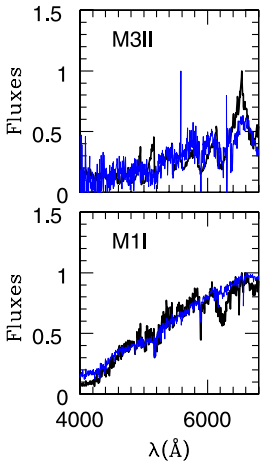

Fig. 9 SDSS-DR2 spectra with the corresponding JHC reference library spectra and its spectro-luminosity type (Table 2)

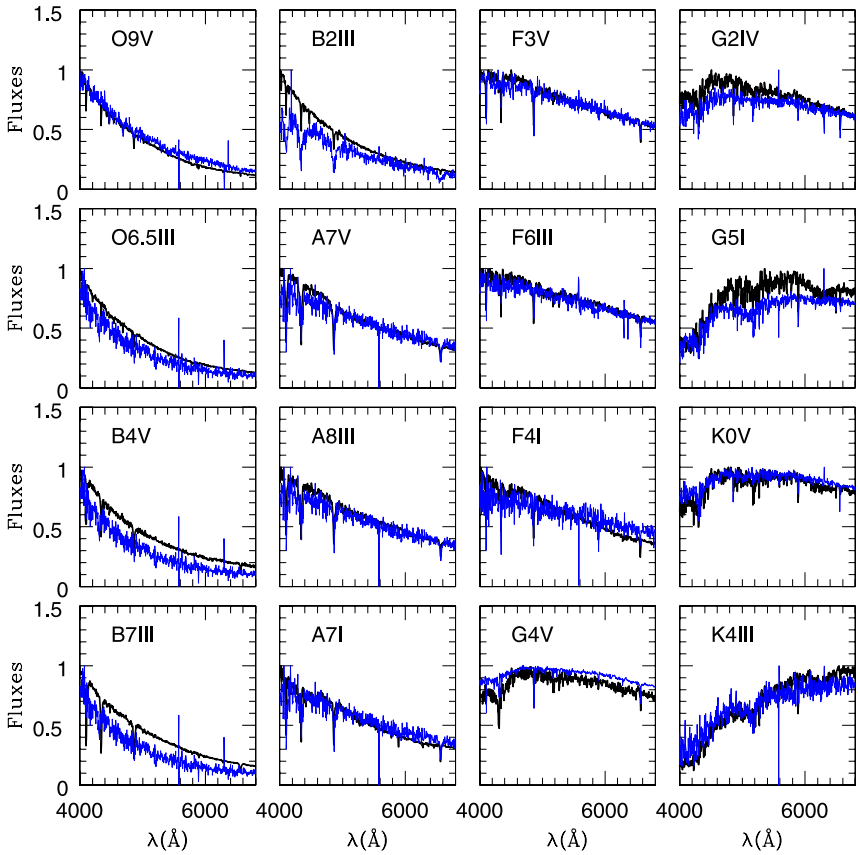

Fig. 10 SDSS-DR2 spectra with the corresponding JHC reference library spectra and its spectro-luminosity type (Table 3)

spectro-luminosity classes of the reference JHC library. The SDSS test data were in fits format with six binary tables attached to the images. We converted the fits images into ASCII files (i.e. two columns of wavelength and fluxes) which are acceptable to the artificial neural networks, using the IRAF task *wspectext*. The spectral resolution of JHC is FWHM = 4.5 Å with one flux value per 1.4 Å and the SDSS has a resolution of about FWHM = 3.25 Å with each flux value at 0.5 Å. The spectral resolution of testing data set is brought to 4.5 Å with sampling rate of 5 Å, thus the final test and train sets are re-binned at 5 Å steps resulting in 561 data points for each spectra in training and test datasets. This is achieved by using appropriate convolution and spline fitting routines. Also both the libraries are normalized into unity.

3 Probabilistic neural networks

The Probabilistic Neural Network (PNN) was developed by Specht (1990). This network provides a general solution to pattern classification problems by following an approach developed in statistics, called Bayesian classifiers. This technique is constructed using ideas from classical probability theory, such as Bayesian classification and classical estimators for probability density functions (pdf) (Parzen 1962), to form a neural network for pattern classification and estimation of class membership (Specht and Romsdahl 1994). The Fig. 1 shows the architecture of PNN. It is a four layer feedforward network consisting of input layer, pattern layer, summation layer and the output layer. The input layer contains n nodes to accept an n-dimensional feature vector ($n = 561$) and is fully connected to the pattern layer which passes the input into pattern layer. The pattern layer consists of k groups of pattern nodes. The kth group in the pattern layer contains N_k number of pattern nodes, where, k is the number of training patterns or classes (i.e. 158 JHC library). The summation layer is having k nodes, one node for each class in the pattern layer. Pattern nodes of each kth group in the pattern layer are connected to the corresponding kth summation node in the summation layer. The probabilistic neural network uses the following estimator for the probability density function of kth group:

$$S_k(X) = \frac{1}{(2\pi\sigma^2)^{n/2}} \frac{1}{N_k} \sum_{i=1}^{N_k} \exp\left(-\frac{\| X - X_{k,i} \|^2}{2\sigma^2}\right), \quad (1)$$

where $X_{k,i} \in \Re^n$ is the center of the kernel, and σ also known as the spread or smoothing parameter which is the deviation of the Gaussian function. Finally at the output

Table 2 List of spectra appearing in the Figs. 6–9

SDSS spectra	Spectro-luminosity class	χ^2 value
spSpec-52294-0327-328	F8III	0.00202269
spSpec-51789-0352-129	F5II	0.00080027
spSpec-51792-0354-184	F0I	0.00086225
spSpec-51943-0300-544	F2I	0.00091199
spSpec-51780-0351-300	F3I	0.00235900
spSpec-51957-0304-551	F4I	0.00157235
spSpec-52017-0366-297	F5I	0.00569906
spSpec-51699-0349-570	F8I	0.00085424
spSpec-51792-0354-388	G0V	0.00106892
spSpec-51957-0304-589	G1V	0.00119590
spSpec-51816-0360-178	G2V	0.00070487
spSpec-51694-0338-614	G3V	0.00074789
spSpec-51913-0274-130	G4V	0.00097247
spSpec-51699-0349-117	G6V	0.00077054
spSpec-51699-0349-252	G7V	0.00065785
spSpec-51699-0349-139	G9V	0.00247349
spSpec-51816-0360-287	G2IV	0.00258696
spSpec-51816-0360-351	G5IV	0.00152954
spSpec-51816-0360-568	G0III	0.00128494
spSpec-51780-0351-608	G2III	0.00176844
spSpec-51694-0338-196	G4III	0.00282095
spSpec-51994-0309-522	G5III	0.00172599
spSpec-51821-0359-556	G6III	0.00141486
spSpec-51699-0349-585	G7III	0.00202173
spSpec-51699-0349-441	G8III	0.00222946

Table 2 (*Continued*)

SDSS spectra	Spectro-luminosity class	χ^2 value
spSpec-51689-0312-285	G9III	0.00181964
spSpec-51788-0355-169	G8II	0.00260567
spSpec-51816-0360-100	G9II	0.00265346
spSpec-51816-0360-497	G0I	0.00359867
spSpec-51673-0313-125	G1I	0.00501194
spSpec-51910-0269-313	G3I	0.00302347
spSpec-51699-0349-340	G5I	0.00416412
spSpec-51699-0349-328	K0V	0.00130288
spSpec-51699-0349-522	K4V	0.00200172
spSpec-51691-0342-037	K5V	0.00259285
spSpec-51780-0351-109	K0III	0.00156690
spSpec-51658-0282-282	K2III	0.00281755
spSpec-51780-0351-309	K3III	0.00319767
spSpec-51955-0298-293	K4III	0.00154753
spSpec-52313-0333-323	K7III	0.00637097
spSpec-51663-0315-540	K6II	0.01685844
spSpec-51821-0359-591	K0I	0.00183392
spSpec-51997-0337-129	K5I	0.00271351
spSpec-51959-0283-455	M0V	0.00648934
spSpec-51989-0363-264	M1V	0.00333934
spSpec-51662-0308-320	M5V	0.01469228
spSpec-52000-0288-064	M3III	0.01214894
spSpec-51658-0282-370	M4III	0.01844951
spSpec-51692-0339-064	M3II	0.01805566
spSpec-52375-0326-135	M1I	0.00607657

layer the pattern vector X will be classified as a class which corresponds to the summation unit with maximum value,

$$C(X) = \arg \max_{1 \le k \le K} (S_k). \tag{2}$$

4 Classification

The application of neural networks to classification problems is conceptually the most consistent with their structure and function. Considering a finite set of states or classes, the objective in classification applications are the assignment of a random samples to one of those states with minimum probability of errors. Each sample is described by a set of parameters which form a vector, usually refereed to as the feature vector. The development of such a classification system can be achieved by appropriately training a neural network in such a way that it provides an output corresponding to one of the classes, provided that, the training sample used in forming its inputs belong to this class. The ability of the neural network to correctly classify a test sample that is close in some sense to one of the training samples, related directly to its generalization ability.

Stellar spectra classification is one of the application areas of the artificial neural networks. Here we use PNN for the classification, an excellent classifier with very fast training process and no local minima issues which outperforms other classifiers including back-propagation. After preprocessing both the training and test datasets, the training set with 158 spectra and 561 flux bins each, are given to the network for the training. So in the input layer of the network there will be 561 number of nodes corresponding to dimension of the spectra. Then the input will be passed to the next layer i.e. pattern layer. All the 158 training samples will be stored in the pattern layer and this layer organizes as groups and each group will be dedicated to one class of spectra. Hence there will be 158 groups in the pattern layer, each containing as many neurons as the number of flux bins in the spectra i.e. 561. In the testing stage, 4999 test spectra

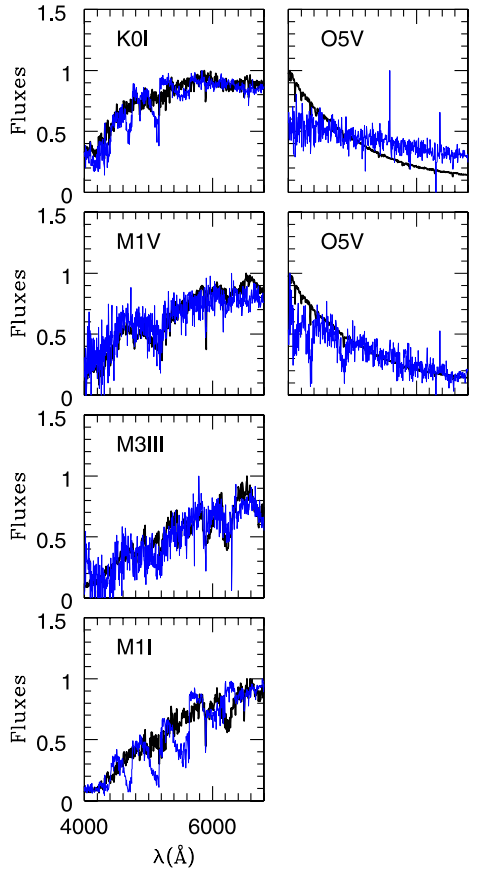

Fig. 11 SDSS-DR2 spectra with the corresponding JHC reference library spectra and its spectro-luminosity type (Table 3)

Table 3 List of spectra with high χ^2 values

SDSS spectra	Spectro-luminosity class	Chisq. value
spSpec-52017-0366-029	O9V	0.00392584
spSpec-52375-0326-557	O6.5III	0.01137241
spSpec-51957-0304-067	B4V	0.02226783
spSpec-52375-0326-565	B7III	0.01909759
spSpec-52000-0288-102	B2III	0.02392821
spSpec-52368-0331-206	A7V	0.00735897
spSpec-51957-0273-141	A8III	0.01864975
spSpec-51789-0352-044	A7I	0.01932068
spSpec-51662-0308-479	F3V	0.00255453
spSpec-52056-0329-587	F6III	0.00313202
spSpec-51613-0305-192	F4I	0.01137626
spSpec-51994-0309-410	G4V	0.00792763
spSpec-52017-0366-310	G2IV	0.01030766
spSpec-52017-0366-293	G5I	0.01894512
spSpec-51818-0358-528	K0V	0.00434149
spSpec-51673-0313-021	K4III	0.01364253
spSpec-51959-0283-365	K0I	0.00883682
spSpec-51928-0291-599	M1V	0.01796854
spSpec-51633-0268-613	M3III	0.01919349
spSpec-51942-0301-308	M1I	0.02038125
spSpec-51999-0336-405	O5V	0.83020251
spSpec-51816-0360-166	O5V	0.02503208

from SDSS will be given to the input layer and the distance between them and the training input vectors will be evaluated and then a vector will be produced by which closeness and similarity of the test data and training data can be judged. Finally these vectors will be summed up in the summation layer to produce an output as vector of probabilities and the maximum probability will be selected at the output node. The neuron corresponding to the highest probability value will be the closest class to the test spectra, for example if neuron S_k in Fig. 1 is having the highest output, it indicates that the given input belongs to the kth group or class.

5 Result

Automated classification of the SDSS-DR2 set of 4999 spectra were carried out using the PNN algorithm. Since we believe that this is a first attempt of classifying the SDSS spectra of originally unknown spectro-luminosity classes into a known set of classes of a JHC reference library; we have provided the complete classification result in Table 4.[1] This table gives the name of star, Spectro-Luminosity class given by ANN and Chi-square error of SDSS and its corresponding JHC reference spectra.

To illustrate the quality of fits of the SDSS spectra with respect to the JHC spectra, we have plotted some typical spectro-luminosity classes from each main spectral types in Figs. 2–5 and 6–9. These plots consist of 16 panels per page covering most of the O to M main spectral types. The JHC reference spectrum are shown in bold black and the SDSS spectra in thin blue lines. Tables 1 and 2 lists the SDSS spectra name; corresponding JHC class obtained by the PNN and the corresponding χ^2 value of these plots.

Figures 10 and 11 show a set of SDSS spectra which are having higher χ^2 values as compared to those of Figs. 2–9; most of which do not fit well with the corresponding JHC spectra obtained by the PNN. Table 3 lists those set of spectra with high χ^2 values.

[1]Table 4 is available for electronic download in the on-line version of this paper.

6 Discussion

As seen from classification result the high value of χ^2 associated with the O5V spectro-luminosity class, two samples of which are shown in Fig. 11. Though there is one spectra in test dataset which is correctly classified into this spectral type and that is spSpec-51673-0313-519 with χ^2 value of 0.00029920 and is shown in Fig. 2 which is well matching with its corresponding JHC spectra; in Figs. 10 and 11 we show the worst case of each class with highest χ^2 values.

A look at the panels in Figs. 2–5 and 6–9 indicates that, except for some spectra, most of them show excellent matches. If one considers a χ^2 value of 0.02 as a kind of limit, then, there are about 600 spectra which has χ^2 value worse than this limit which corresponds to a success rate of about 88%.

We could classify the stars based on their temperature and luminosity classes with PNN technique, and only a few seconds were required to classify all 4999 spectra (in the test session). Considering this high speed of classification, it would enable us to apply this technique to classify larger datasets of the latest SDSS data releases. We are hoping to achieve better results in the future classification of SDSS datasets by using training samples from SDSS itself (i.e. the sample can be taken from our classification results by selecting the best match spectra of each spectral and luminosity class having smallest χ^2 value in this catalog).

This work also encourages to use ANN based techniques to classify the complete SDSS set of spectra in near future.

Acknowledgements The first author thanks IUCAA and IASBS for providing the computational facilities for this work.

References

Abazajian, K., et al.: Astron. J. **126**, 2081 (2003)
Abazajian, K., et al.: Astron. J. **128**, 502 (2004)
Abazajian, K., et al.: Astron. J. **129**, 1755 (2005)
Adelman-McCarthy, J., et al.: Astrophys. J. Suppl. Ser. **162**, 38 (2006)
Adelman-McCarthy, J., et al.: Astrophys. J. Suppl. Ser. **172**, 634 (2007a)
Adelman-McCarthy, J., et al.: Astrophys. J. Suppl. Ser. (2007b, submitte)
Allende Prieto, C.: Astron. Nachr. **325**(6–8), 604 (2004)
Ball, N.M., Loveday, J., Fukugita, M., Nakamura, O., Okamura, S., Brinkmann, J., Brunner, R.J.: Mon. Not. R. Astron. Soc. **348**, 1038 (2004)
Bazarghan, M.: Bull. Astr. Soc. India **36**, 1–54 (2008)
Gulati, R.K., Gupta, R., Khobragade, S.: Astrophys. J. **426**, 340 (1994a)
Gulati, R.K., Gupta, R., Khobragade, S.: Vistas Astron. **38**, 293 (1994b)
Gulati, R.K., Gupta, R., Khobragade, S.: In: Shaw, R.A., Payne, H.E., Hayes, J.J.E. (eds.) Astronomica Data Analysis Software and Systems IV. ASP Conf. Ser., vol. 77, p. 253. (1995)
Gupta Ranjan, Singh Harinder, P., Volk, K., Kwok, S.: Astrophys. J. Suppl. Ser. **152**, 201 (2004)
Jacoby, G.H., Hunter, D.A., Christian, C.A.: Astrophys. J. Suppl. Ser. **56**, 257 (1984)
Parzen, E.: Ann. Math. Stat. **3**, 1065 (1962)
Qin, D.-M., Guo, P., Hu, Z.-Y., Zhao, Y.-H.: Chin. J. Astron. Astrophys. **3**, 277 (2003)
Singh Harinder, P., Gulati Ravi, K., Gupta Ranjan: Mon. Not. R. Astron. Soc. **295**, 312 (1998)
Singh Harinder, P., Gupta, R.: Large telescopes and virtual observatory: visions for the future. In: 25th Meeting of the IAU. Joint Discussion, vol. 8 (2003)
Specht, D.F.: Neural Netw. **1**(3), 109 (1990)
Specht, D.F., Romsdahl, H.: In: Proceedings of the IEEE International Conference on Neural Networks, vol. 2, p. 1203 (1994)
Stoughton, C.: Astron. J. **123**, 3487 (2002)
Vanzella, E., et al.: Astron. Astrophys. **423**, 761 (2004)
Von Hipple, T., Storrie-Lombardi, L.J., Storrie-Lombardi, M.C., Irwin, M.J.: Mon. Not. R. Astron. Soc. **269**, 97 (1994)
Weaver, W.B., Torres-Dodgen, A.V.: Astrophys. J. **446**, 300 (1995)
York, D.G., Adelman, J., Anderson, J.E., et al.: Astron. J. **120**, 1579 (2000)