Timbre Identification - A Seminar Report **MUS-17**

Carina Schäffer

July 5, 2017

Introduction

As defined within the Acoustical Society of America (ASA) Standard Term Database, timbre is "that multidimensional attribute of auditory sensation which enables a listener to judge that two non-identical sounds, similarly presented and having the same loudness, pitch, spatial location, and duration, are dissimilar"[1]. Alternatively, the Cambridge dictionary gives an easier to understand definition: "[Timbre is] a quality of sound that makes voices or musical instruments sound different from each other" [2]. For the term "timbre identification", this means we are striving to computationally identify different timbres by preselecting certain features, e.g. instruments or musicians (as done so in [3]), in order to clearly distinguish one preselected parameter from the other. This is a machine learning problem, used amongst others for genre categorization, automatic score creation or track separation.

History

The foundations of timbre identification lie in the late 1970s, when John Grey from Stanford University started initial investigations on the task of musical instrument identification [4]. In 1999, Marques and Moreno, Cambridge Research Laboratory, used Support Vector Machines (SVMs) to reach an 70% accuracy for classification of eight different musical instruments [5]. Just one year later, Fujinaga and MacMillan trained an k-Nearest Neighbor (k-NN) system to reach an accuracy of 68% on a much larger data set of 23 different instruments at John Hopkins University, Baltimore [6]. In later years, those two systems became the most prominent for musical instrument classification and were developed further to reach higher accuracies. Essid, Richard and David were able to get results of 87% accuracy with SVMs at Paris-Saclay in 2006 [7], while Kaminskyj and Czaszejko at Monash University, Melbourne, even reached accuracy scores as high as 93% on instrument classification using a k-NN system [8].

Classification of Musical Timbre Using Bayesian Networks

Having started my literature search from Guo et al. [3], in this report I want to present the paper "Classification of Musical Timbre Using Bayesian Networks" by Patrick J. Donnelly and John W. Sheppard, published in 2014 [9]. It introduces a concept that was unto then quite unknown within the field of musical instrument classification: Bayesian networks. Using a data set that not only covers a wide range of different instruments but is also openly available, Donnelly compares already commonly used methods, namely the support vector machine algorithm and the k-nearest neighbor algorithm, with a set of different Bayesian network models which he proposes within the course of his paper. Later on, he conducts a couple of different experiments, determining the classification accuracy on instruments. Donnelly is able to show that, using Bayesian networks, his classifications show promising results, with some outperforming the older approaches. As the paper only explores monophonic single instrument classification, the task of identification and classification of polyphonic data remains unapproached by Donnelly. [10] suggests though that the author is further continuing with his research on musical instrument classification using timbre segmentation and Bayesian networks.

Algorithms

The different classifiers, that Donnelly and Sheppard compare with each other, are based on three different concepts, that I want to introduce here:

k-Nearest Neighbor

The k-Nearest Neighbor algorithm, or short k-NN algorithm, will classify a previously unknown example with the most common class found amongst the example's nearest neighbors. The nearest neighbor is evaluated with a distance metric, most common here is the Euclidean distance

$$D(\boldsymbol{u}, \boldsymbol{v}) = \sqrt{\sum_{i=1}^{n} (\boldsymbol{u}_i - \boldsymbol{v}_i)^2},$$

where $\boldsymbol{u}$ and $\boldsymbol{v}$ are n-dimensional feature vectors. For the classification of musical instruments, the k-NN algorithm proceeds as follows: First, each sample in the test set is compared to a subset of examples from the training set using the distance metric. It is then assigned with the most common class label among its k nearest neighbors.

Support Vector Machine

The support vector machine (SVM) algorithm is a discriminant-based method for classification or regression. It constructs a hyperplane in high dimensional space that represents the largest margin separating two classes of data. In multiclass problems, the SVM algorithm produces "one-versus-all" binary classifiers that try to separate each class from all other possible classes. If the kernel function of the feature vector is the feature vector itself, the support vector machine will take the form of a linear classifier. Otherwise, if the kernel is a non-linear function, the features are projected into higher-order space. Thus, the algorithm can fit the maximum margin hyperplane in the transformed feature space which again allows for clear separation of different classes.

Bayesian Networks

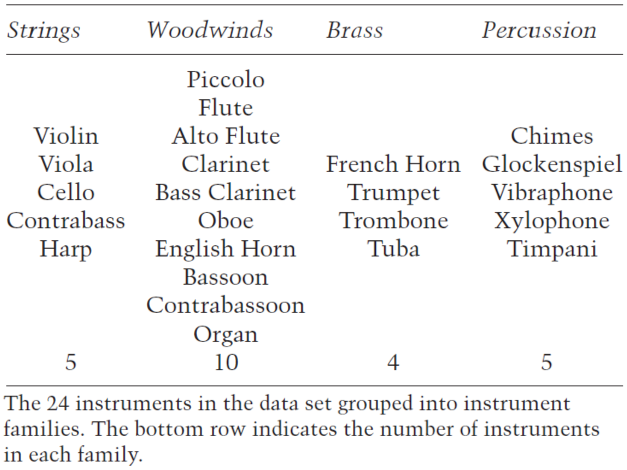
Bayesian networks are probabilistic graph models composed of random variables that are represented as nodes, and their conditional dependencies, represented as directed edges. The joint probability of multiple represented variables is the product of the individual probabilities of each variable, conditioned on the node's parent variables. The Bayesian classifier is then defined as:

$$classify(\boldsymbol{f}) = \underset{c \in C}{argmax} P(c) \prod_{f \in \boldsymbol{f}} P(f|parent(f)),$$

where $P(c)$ is the prior probability of class c and $P(f|parent(f))$ the conditional probability of feature f given the values of the variable's parents. The Bayesian classifier finds the class label with the highest probability of explaining the values of the feature vector.

Feature Extraction

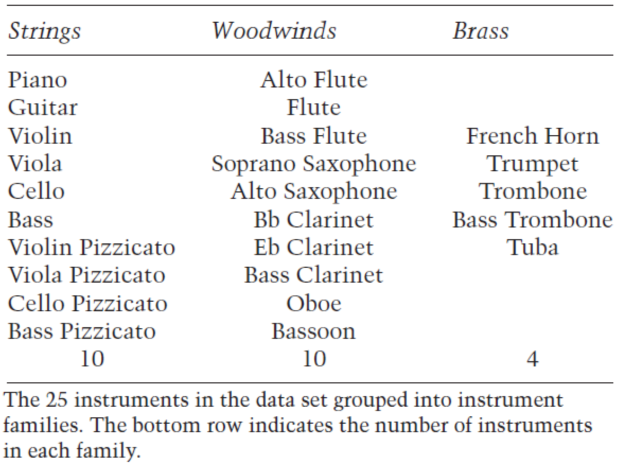
For comparing the different algorithms, two different data sets are used: the EastWest data set [11] (see Figure 1) and the Iowa data set [12] (see Figure 2). The Iowa data set was created by the Electronic Music Studios at the University of Iowa in 2013 and modified to get individual files with each containing a single note, thereby creating a data set containing 4,521 samples over a total of 25 musical instruments. On the contrary, the EastWest data set was specifically created for the task, covering 1000 audio examples for each of the 24 recorded instruments. Within the audio files, you can hear an instrument sustaining a single note for 1s. Each file is 2s long to include the attack and the decay of the note. In order to be able to mathematically express the



Strings	*Woodwinds*	*Brass*	*Percussion*
	Piccolo		
	Flute		
Violin	Alto Flute		Chimes
Viola	Clarinet	French Horn	Glockenspiel
Cello	Bass Clarinet	Trumpet	Vibraphone
Contrabass	Oboe	Trombone	Xylophone
Harp	English Horn	Tuba	Timpani
	Bassoon		
	Contrabassoon		
	Organ		
5	10	4	5

The 24 instruments in the data set grouped into instrument families. The bottom row indicates the number of instruments in each family.

Figure 1: EastWest Data Set of Instruments



Strings	*Woodwinds*	*Brass*
Piano	Alto Flute	
Guitar	Flute	
Violin	Bass Flute	French Horn
Viola	Soprano Saxophone	Trumpet
Cello	Alto Saxophone	Trombone
Bass	Bb Clarinet	Bass Trombone
Violin Pizzicato	Eb Clarinet	Tuba
Viola Pizzicato	Bass Clarinet	
Cello Pizzicato	Oboe	
Bass Pizzicato	Bassoon	
10	10	4

The 25 instruments in the data set grouped into instrument families. The bottom row indicates the number of instruments in each family.

Figure 2: Iowa Data Set of Instruments

different recorded notes in the way that the formerly introduced algorithms will be able to handle them, the audio files are transformed to small vectors of relevant numeric features. As has recently come to attention, the choice of those relevant features heavily influences the outcome of the chosen learning algorithms used for classification [13]. Using fast Fourier transform over 20 100ms-slots, the amplitude is then obtained as a function of frequencies. These frequencies are then grouped into ten exponentially increasing windows on a range from 0 to 22,050Hz with each window having twice the size of the previous one. Afterwards, for each frequency window the peak amplitude is extracted as feature.

Bayesian Network Models

After having introduced the concept of Bayesian networks earlier, let's now take a look at the precise Bayesian classifiers used for the comparison:

- The naive Bayes (NB) classifier assumes, that all evidence nodes are conditionally independent of each other given the class:

  $$P(c|\boldsymbol{f}) = P(c) \cdot \prod_{f \in \boldsymbol{f}} P(f|c)$$

  It is chosen here as a baseline Bayesian model.

- In the following abbreviated as BN-F, the next model takes frequency dependencies into account. Each frequency feature is assumed to be conditionally dependent on the previous frequency feature within a single time window:

  $$\begin{aligned} P(c|\boldsymbol{f}) =& P(c) \cdot \prod_{i=1}^{20} P(f_1^i|c) \\ & \cdot \left(\prod_{i=1}^{20} \prod_{j=2}^{10} P(f_1^i|f_{j-1}^i, c) \right) \end{aligned}$$

- The third model, called BN-T, considers time dependencies, containing conditional dependencies in the time domain but not in the frequency domain:

  $$\begin{aligned} P(c|\boldsymbol{f}) =& P(c) \cdot \prod_{j=1}^{10} P(f_j^1|c) \\ & \cdot \left(\prod_{i=2}^{20} \prod_{j=1}^{10} P(f_j^i|f_j^{i-1}, c) \right) \end{aligned}$$

- The last model to be introduced both considers frequency and time dependencies and shall be called BN-FT:

$$\begin{aligned} P(c|\boldsymbol{f}) =& P(c) \cdot P(f_1^1|c) \\ & \cdot \prod_{i=2}^{20} P(f_1^i|f_1^{i-1}, c) \\ & \cdot \prod_{i=2}^{10} P(f_j^1|f_{j-1}^1, c) \\ & \cdot \left(\prod_{i=2}^{20} \prod_{j=2}^{10} P(f_j^i|f_j^{i-1}, f_{j-1}^i, c) \right) \end{aligned}$$

Experiments

There were four experiments conducted to compare the different models with each other:

1. Instrument and family identification,
2. Instrument Identification within Family,
3. Classification Accuracy by Data Set Size, and
4. Repetition of Experiments 1 and 2 with Iowa Data Set.

Results

The results of Experiment 1 can be seen in Figure 3 (Accuracy), Figure 4 (Statistical Significance) and Figure 5 (confusion Matrices). One can clearly see, that all the Bayesian network models (with the excemption of the naive Bayes classifier) outperformed the k-NN and SVM algorithms on instrument classification. FOr the sake of the difference in linear and non-linear kernels in SVM algorithms, both a SVM with a linear kernel (SVM-L) and one with a polynomial kernel (SVM-Q) are used for comparing. The Bayesian models perform worse on family identification though. There seems to be an increased confusion between brass and woodwind instruments, compared to string or percussion instruments. SVMs, k-NN and naive Bayes have a higher confusion between strings and either brass or woodwind, though.

In Figure 7b, the classification accuracy in Experiment 2 is shown. All Bayesian classifiers (again except naive Bayes) reach more than 99% accuracy for all families but woodwinds. Nearly all algorithms achieve the highest accuracies for percussion instruments.

Experiment 3 was conducted as to be able to better understand the influence of the data set size n the performance of the different models. Varying the size from 100 to 1,000 samples in increments of 100 for each instrument, the Bayesian models interestingly reach the highest accuracies for a set size of 500 to 800. More predictable, SVMs and k-NN constantly improve with an increasing number of samples. Also of note is, that Bayesian models were able to achieve much higher accuracies with far fewer examples than either SVMs or k-NN could.

As the classifiers were both trained and tested on the EastWest data set, one last Experiment was conducted to test their performance on the yet unknown Iowa data set. Although it is a significantly smaller data set, the results are consistent with the previous ones considering the same data size. The results of Experiment 4 (Accuracy, statistical significance and confusion matrices) are shown in Figures 7a, 6 and 7c.

Conclusion

Within this report, I have first given an introduction to timbre identification. After a short excursion into the history of the matter, I have presented the most prominent algorithms in this field and given a short overview about the topic of feature extraction. Afterwards, I described four different Bayesian classifiers which were then used to compare the success of Bayesian networks versus k-NN and SVM algorithms with the result that they (apart from the naive Bayes classifier) fared quite well, in particular when taking both time and frequency dependencies into account.

References

[1] Acoustic Society of America. *timbre - Welcome to ASA Standards.* URL: `http://asastandards.org/Terms/timbre/`.

[2] Cambridge University Press. *timbre Bedeutung im Cambridge Englisch Wörterbuch.* URL: `http://dictionary.cambridge.org/de/worterbuch/englisch/timbre`.

[3] Jinxi Guo et al. "Timbre Identification of Instrumental Music via Energy Distribution Modeling". In: *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service.* ICIMCS '15. Zhangjiajie, Hunan, China: ACM, 2015, 67:1–67:5.

[4] John M Grey. "Multidimensional perceptual scaling of musical timbres". In: *the Journal of the Acoustical Society of America* 61.5 (1977), pp. 1270–1277.

[5] Janet Marques and Pedro J Moreno. "A study of musical instrument classification using gaussian mixture models and support vector machines". In: *Cambridge Research Laboratory Technical Report Series CRL* 4 (1999).

[6] Ichiro Fujinaga and Karl MacMillan. "Realtime Recognition of Orchestral Instruments." In: *ICMC.* 2000.

[7] Slim Essid, Gaël Richard, and Bertrand David. "Musical instrument recognition by pairwise classification strategies". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.4 (2006), pp. 1401–1412.

[8] Ian Kaminskyj and Tadeusz Czaszejko. "Automatic recognition of isolated monophonic musical instrument sounds using kNNC". In: *Journal of Intelligent Information Systems* 24.2 (2005), pp. 199–221.

[9] Patrick J. Donnelly and John W. Sheppard. "Classification of Musical Timbre Using Bayesian Networks". In: *Comput. Music J.* 37.4 (Dec. 2014), pp. 70–86.

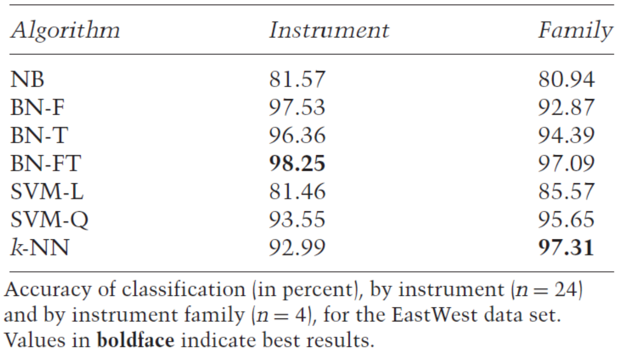
[10] Patrick Joseph Donnelly. "Bayesian Approaches to Musical Instrument Classification using Timbre Segmentation". PhD thesis. MONTANA STATE UNIVERSITY Bozeman, 2012.

[11] Numerical Intelligent Systems Laboratory. *Index of /instruments.* http://nisl.cs.montana.edu/instruments. Accessed on May 31st, 2017.

[12] University of Iowa Electronic Music Studios. *Musical Instrument Samples.* http://theremin.music.uiowa.edu/MIS.html. Accessed on May 31st, 2017.

[13] Jeremiah D Deng, Christian Simmermacher, and Stephen Cranefield. "A study on feature analysis for musical instrument classification". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 38.2 (2008), pp. 429–438.

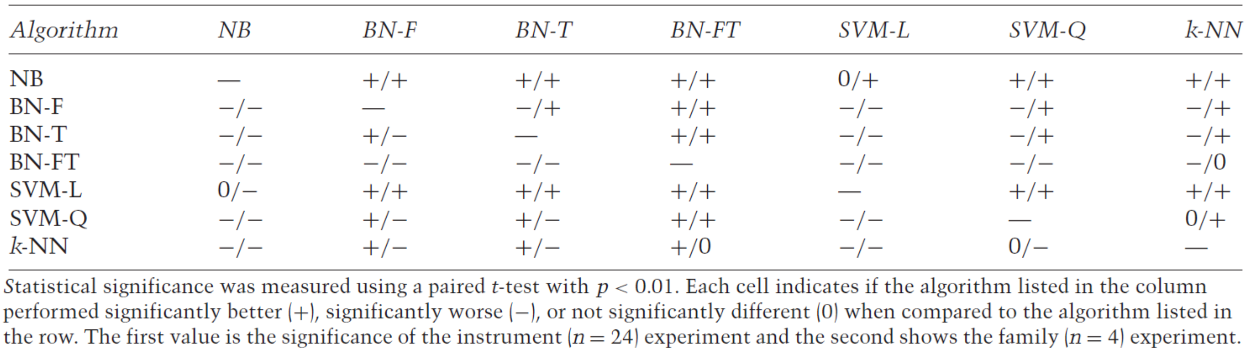
Appendix



Algorithm	*Instrument*	*Family*
NB	81.57	80.94
BN-F	97.53	92.87
BN-T	96.36	94.39
BN-FT	**98.25**	97.09
SVM-L	81.46	85.57
SVM-Q	93.55	95.65
k-NN	92.99	**97.31**

Accuracy of classification (in percent), by instrument ($n = 24$) and by instrument family ($n = 4$), for the EastWest data set. Values in **boldface** indicate best results.

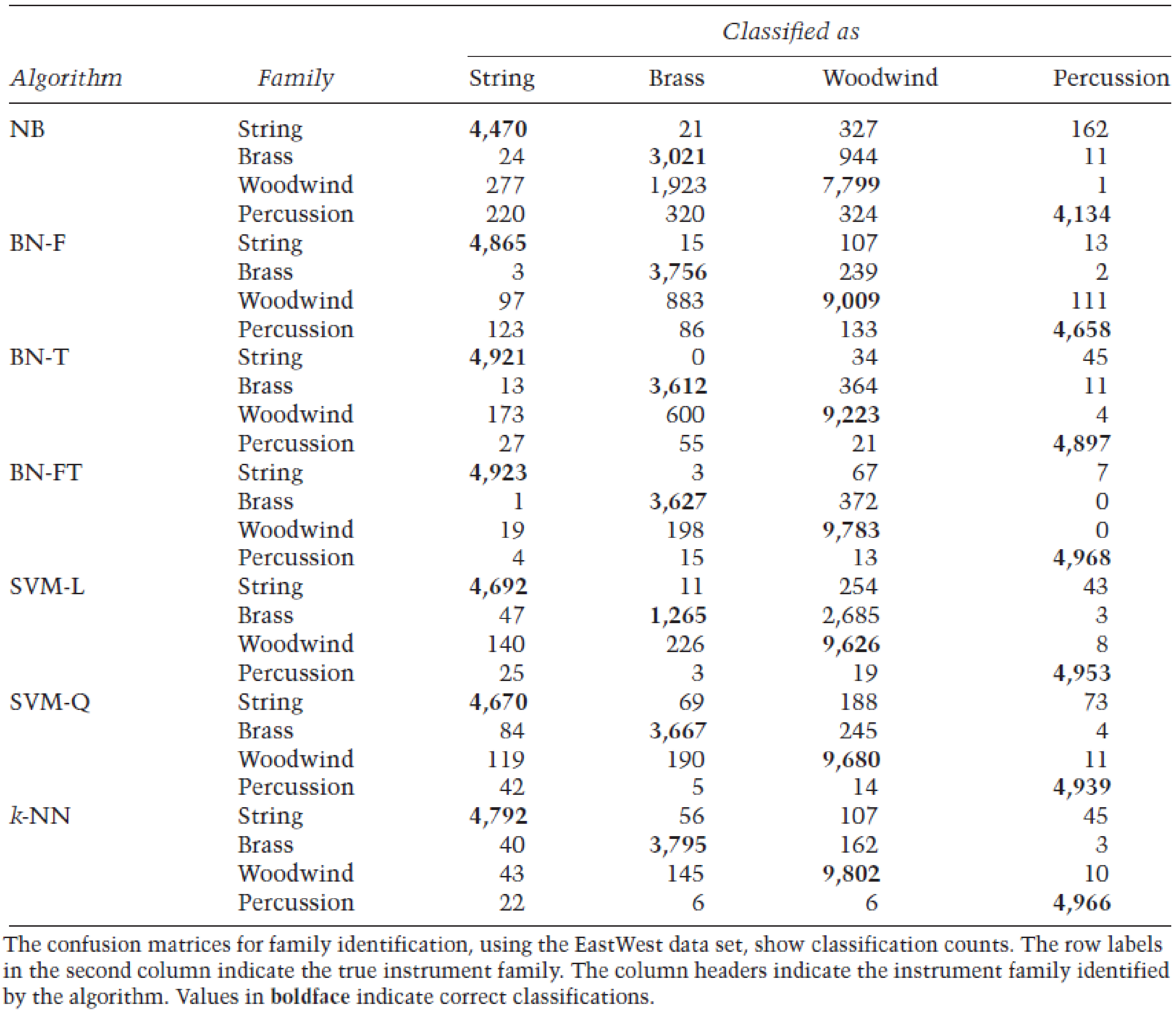
Figure 3: Classification Accuracy in Experiment 1



Algorithm	*NB*	*BN-F*	*BN-T*	*BN-FT*	*SVM-L*	*SVM-Q*	*k-NN*
NB	—	+/+	+/+	+/+	0/+	+/+	+/+
BN-F	−/−	—	−/+	+/+	−/−	−/+	−/+
BN-T	−/−	+/−	—	+/+	−/−	−/+	−/+
BN-FT	−/−	−/−	−/−	—	−/−	−/−	−/0
SVM-L	0/−	+/+	+/+	+/+	—	+/+	+/+
SVM-Q	−/−	+/−	+/−	+/+	−/−	—	0/+
k-NN	−/−	+/−	+/−	+/0	−/−	0/−	—

Statistical significance was measured using a paired *t*-test with $p < 0.01$. Each cell indicates if the algorithm listed in the column performed significantly better (+), significantly worse (−), or not significantly different (0) when compared to the algorithm listed in the row. The first value is the significance of the instrument ($n = 24$) experiment and the second shows the family ($n = 4$) experiment.

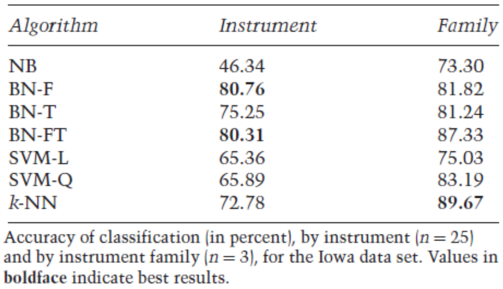
Figure 4: Statistical Significance from Experiment 1



		Classified as			
Algorithm	*Family*	String	Brass	Woodwind	Percussion
NB	String	**4,470**	21	327	162
	Brass	24	**3,021**	944	11
	Woodwind	277	1,923	**7,799**	1
	Percussion	220	320	324	**4,134**
BN-F	String	**4,865**	15	107	13
	Brass	3	**3,756**	239	2
	Woodwind	97	883	**9,009**	111
	Percussion	123	86	133	**4,658**
BN-T	String	**4,921**	0	34	45
	Brass	13	**3,612**	364	11
	Woodwind	173	600	**9,223**	4
	Percussion	27	55	21	**4,897**
BN-FT	String	**4,923**	3	67	7
	Brass	1	**3,627**	372	0
	Woodwind	19	198	**9,783**	0
	Percussion	4	15	13	**4,968**
SVM-L	String	**4,692**	11	254	43
	Brass	47	**1,265**	2,685	3
	Woodwind	140	226	**9,626**	8
	Percussion	25	3	19	**4,953**
SVM-Q	String	**4,670**	69	188	73
	Brass	84	**3,667**	245	4
	Woodwind	119	190	**9,680**	11
	Percussion	42	5	14	**4,939**
k-NN	String	**4,792**	56	107	45
	Brass	40	**3,795**	162	3
	Woodwind	43	145	**9,802**	10
	Percussion	22	6	6	**4,966**

The confusion matrices for family identification, using the EastWest data set, show classification counts. The row labels in the second column indicate the true instrument family. The column headers indicate the instrument family identified by the algorithm. Values in **boldface** indicate correct classifications.

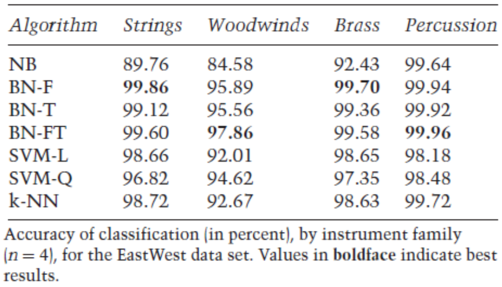
Figure 5: Confusion Matrices for Experiment 1



Algorithm	*Instrument*	*Family*
NB	46.34	73.30
BN-F	**80.76**	81.82
BN-T	75.25	81.24
BN-FT	**80.31**	87.33
SVM-L	65.36	75.03
SVM-Q	65.89	83.19
k-NN	72.78	**89.67**

Accuracy of classification (in percent), by instrument ($n = 25$) and by instrument family ($n = 3$), for the Iowa data set. Values in **boldface** indicate best results.

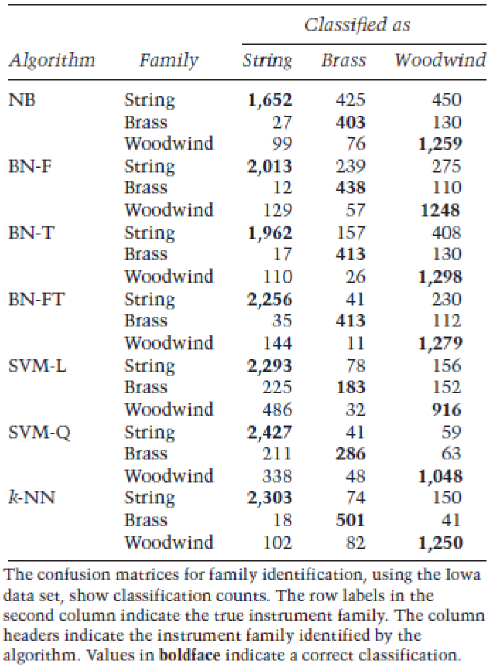
(a) Classification Accuracy in Experiment 4



Algorithm	*Strings*	*Woodwinds*	*Brass*	*Percussion*
NB	89.76	84.58	92.43	99.64
BN-F	**99.86**	95.89	**99.70**	99.94
BN-T	99.12	95.56	99.36	99.92
BN-FT	99.60	**97.86**	99.58	**99.96**
SVM-L	98.66	92.01	98.65	98.18
SVM-Q	96.82	94.62	97.35	98.48
k-NN	98.72	92.67	98.63	99.72

Accuracy of classification (in percent), by instrument family ($n = 4$), for the EastWest data set. Values in **boldface** indicate best results.

(b) Classification Accuracy in Experiment 2

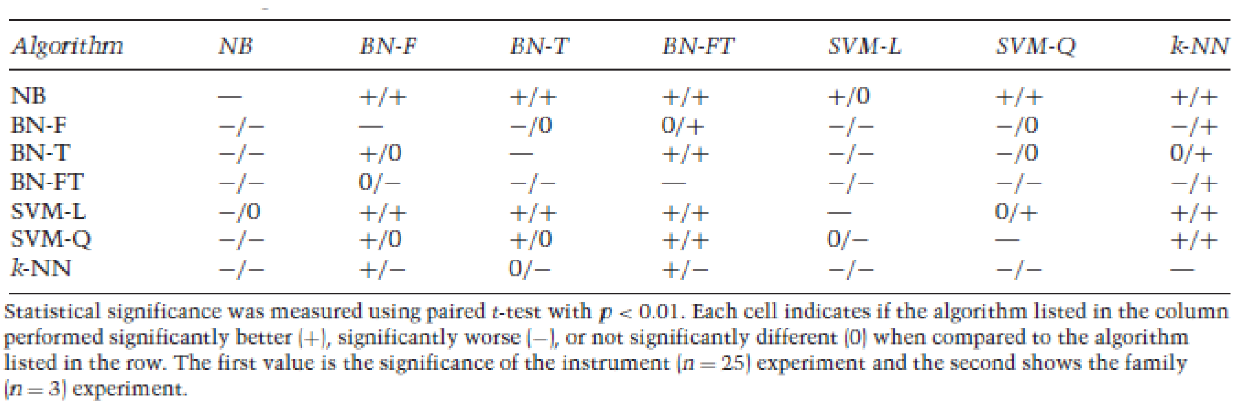


		Classified as		
Algorithm	*Family*	*String*	*Brass*	*Woodwind*
NB	String	**1,652**	425	450
	Brass	27	**403**	130
	Woodwind	99	76	**1,259**
BN-F	String	**2,013**	239	275
	Brass	12	**438**	110
	Woodwind	129	57	**1248**
BN-T	String	**1,962**	157	408
	Brass	17	**413**	130
	Woodwind	110	26	**1,298**
BN-FT	String	**2,256**	41	230
	Brass	35	**413**	112
	Woodwind	144	11	**1,279**
SVM-L	String	**2,293**	78	156
	Brass	225	**183**	152
	Woodwind	486	32	**916**
SVM-Q	String	**2,427**	41	59
	Brass	211	**286**	63
	Woodwind	338	48	**1,048**
k-NN	String	**2,303**	74	150
	Brass	18	**501**	41
	Woodwind	102	82	**1,250**

The confusion matrices for family identification, using the Iowa data set, show classification counts. The row labels in the second column indicate the true instrument family. The column headers indicate the instrument family identified by the algorithm. Values in **boldface** indicate a correct classification.

(c) Confusion Matrices for Experiment 4

Figure 7: Pictures of animals



Algorithm	*NB*	*BN-F*	*BN-T*	*BN-FT*	*SVM-L*	*SVM-Q*	*k-NN*
NB	—	+/+	+/+	+/+	+/0	+/+	+/+
BN-F	−/−	—	−/0	0/+	−/−	−/0	−/+
BN-T	−/−	+/0	—	+/+	−/−	−/0	0/+
BN-FT	−/−	0/−	−/−	—	−/−	−/−	−/+
SVM-L	−/0	+/+	+/+	+/+	—	0/+	+/+
SVM-Q	−/−	+/0	+/0	+/+	0/−	—	+/+
k-NN	−/−	+/−	0/−	+/−	−/−	−/−	—

Statistical significance was measured using paired t-test with $p < 0.01$. Each cell indicates if the algorithm listed in the column performed significantly better (+), significantly worse (−), or not significantly different (0) when compared to the algorithm listed in the row. The first value is the significance of the instrument ($n = 25$) experiment and the second shows the family ($n = 3$) experiment.

Figure 6: Statistical Significance from Experiment 4