Pattern Recognition Letters xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Bayes covariant multi-class classification☆

Ondrej Šuch[a,b,*], Santiago Barreda[c]

[a] University of Žilina, Žilina, Slovakia
[b] Mathematical Institute of Slovak Academy of Sciences, Banská Bystrica, Slovakia
[c] Department of Linguistics, University of California, Davis, CA, United States

ARTICLE INFO

Article history:
Received 9 December 2015
Available online xxx

Keywords:
Multi-class classification
Bradley–Terry model
Bayes classifier
Combining binary classifiers
TIMIT
Vowel classification

ABSTRACT

We consider multi-class classification models built from complete sets of pairwise binary classifiers. The Bradley–Terry model is often used to estimate posterior distributions in this setting. We introduce the notion of Bayes covariance, which holds if the multi-class classifier respects multiplicative group action on class priors. As a consequence, a Bayes covariant method yields the same result whether new priors are considered before or after combination of the individual classifiers, which has several practical advantages for systems with feedback. In the paper, we construct a Bayes covariant combining method and compare it with previously published methods in both Monte Carlo simulations as well as on a practical speech frame recognition task.

1. Introduction

Statistical and machine-learning classification methods have found widespread applications in industry, as well as in scientific research. Successful applications include optical character recognition [1], speech recognition systems [2], automated medical diagnoses [3] and credit-risk scoring [4]. Although in some practical applications binary decisions may be sufficient (e.g. cancer/no cancer decision), most applications require correct classification among multiple classes.

Broadly speaking, multi-class classification will pose a more challenging problem than binary classification. One reason for this is that the set of boundaries among multiple classes may be more complex and thus may be harder to learn than the boundary between two classes. Another reason is that several powerful machine learning methods for classification of two classes have no direct analogues for multiple classes, making these methods inapplicable for those faced with a multi-class problem. Important examples of such methods include support vector machines [5,6] and Adaboost [7].

There are many ways to reduce the multi-class classification of K classes to binary classification subproblems. One common approach is one-vs-all classification when one trains K classifiers to distinguish each class from all of the rest [8]. Another common approach is all-vs-all when one trains $\binom{K}{2}$ pairwise classifiers [9]. Other approaches have been proposed based on error correcting coding theory [10–14] and on training statistical meta-classifiers [15].

In our work, we consider the question of combining the output of binary classifiers in an all-vs-all setting. Some reasons to consider this approach rather than the one-vs-all approach [8] include:

- larger number of parameters allow for more powerful models,
- simpler and faster training of individual classifiers compared to one-vs-all ([8, pp. 123–124]),
- when samples are densely packed in Euclidean space, the all-vs-all boundaries should be simpler, and thus easier to learn than one-vs-all boundaries; for example English vowels lie essentially in a 2-dimensional space [16],
- larger number of binary models allows for some tolerance of imprecision of individual classifiers ([17], [8, p. 102], [18]). Imprecise computation is typical for neuromorphic circuits for classification problems, which on the other hand are highly parallel and highly energy efficient [19,20].

Bayes theorem provides a rigorous foundation of classification. The theorem explains the crucial role played by class priors on the outcome of classification (cf. (2)). Usually, class priors are a fixed quantity during classification. However, in multi-tiered systems with feedback, one may desire to reevaluate evidence with different priors based on feedback from other tiers. For instance, a typical automated speech recognition system consists of three parts – an acoustic model, a lexicon and a language model [21].

☆ This paper has been recommended for acceptance by Maria De Marsico.

* Corresponding author at: University of Žilina, Žilina, Slovakia. Fax: +421 415134312.
E-mail address: ondrej.such@fri.uniza.sk, ondrej.such@gmail.com (O. Šuch).

http://dx.doi.org/10.1016/j.patrec.2016.08.014
0167-8655/

Solving $r'_{12} = r'_{23} = r'_{31}$ results in:

$$A^3 = \frac{(\frac{1}{r_{23}} - 1)(\frac{1}{r_{31}} - 1)}{(\frac{1}{r_{12}} - 1)^2}, \qquad B^3 = \frac{(\frac{1}{r_{31}} - 1)^2}{(\frac{1}{r_{12}} - 1)(\frac{1}{r_{23}} - 1)}. \tag{15}$$

From (9) and (11) we conclude that the posterior of the Bayes covariant method M must be:

$$M(\mathbf{R}) = \frac{1}{1 + 1/A + 1/B}(1, \tfrac{1}{A}, \tfrac{1}{B}). \tag{16}$$

Eqs. (15) and (16) define M uniquely. The resulting combining method is clearly 3-symmetric and it remains to check for Bayes covariance. Let $\mathbf{s} > 0$ be another reweighing vector. Then for $\mathbf{R^s} = (r''_{ij})$ we have:

$$\frac{s_2}{s_1}\left(\frac{1}{r_{12}} - 1\right) = \frac{1}{r''_{12}} - 1 \tag{17}$$

$$\frac{s_1}{s_3}\left(\frac{1}{r_{31}} - 1\right) = \frac{1}{r''_{31}} - 1 \tag{18}$$

$$\frac{s_3}{s_2}\left(\frac{1}{r_{23}} - 1\right) = \frac{1}{r''_{23}} - 1. \tag{19}$$

It follows that:

$$A\frac{s_1}{s_2}\left(\frac{1}{r''_{12}} - 1\right) = \frac{s_3}{s_1 B}\left(\frac{1}{r''_{31}} - 1\right) = \frac{s_2}{s_3}\frac{B}{A}\left(\frac{1}{r''_{23}} - 1\right) \tag{20}$$

and thus:

$$M(\mathbf{R^s}) \propto \left(1, \frac{s_2}{As_1}, \frac{s_3}{s_1 B}\right) \propto \left(s_1, \frac{s_2}{A}, \frac{s_3}{B}\right). \tag{21}$$

This concludes verification of Bayes covariance and the proof of the theorem. □

5. A general Bayes covariant combining method

We will now construct Bayes covariant classifiers for cases with more than three categories ($K > 3$). Consider the moduli F of all feasible matrices. Inside F there is a submanifold B of feasible matrices for which (3) is a consistent system, which we shall call the *Bradley–Terry manifold*. Let us denote by P the point on the Bradley–Terry manifold corresponding to $r_{ij} = 1/2$ for $i \neq j$. We will consider only methods M for which

$$M(P) = \left(\frac{1}{K}, \tfrac{1}{K}, \ldots\right). \tag{22}$$

There is a natural action of the group $\mathbf{G}$ of reweighing vectors $\mathbf{q} > 0$ on F given by (6). Since the action is simply transitive on the Bradley–Terry manifold B, it follows from (9) that a Bayes covariant method is uniquely determined on the Bradley–Terry manifold.

Going back to proof of Theorem 1, we see that we took advantage of a set S of matrices satisfying (10). The set contained a single representative of each orbit under the action of reweighing vectors on F and by 3-symmetry we knew the exact value of the combining method on S.

In general, one may desire S to be a manifold of codimension $K - 1$ inside the variety of all feasible matrices and prescribe that

$$M(s) = \left(\frac{1}{K}, \tfrac{1}{K}, \ldots\right) \quad \text{for } s \in S. \tag{23}$$

If we express any feasible point $f \in F$ as $f = s^{\mathbf{q}}$ for reweighting vector $\mathbf{q} > 0$ and $s \in S$, we will have from (9):

$$M(f) \propto \mathbf{q}. \tag{24}$$

Since the group of reweighing vectors $\mathbf{G}$ acts transitively on the Bradley–Terry manifold, S has to have a single point intersection with the Bradley–Terry manifold. To arrive at easily computable expressions, we propose to take for S the set of points Q such that QP is orthogonal to the tangent space of the Bradley–Terry manifold at the point P. This set is dependent on parameterization of the variety of feasible matrices, and may not contain a single representative of each orbit. Since we propose a linearly defined set S, it is natural to consider a parameterization in which the group action is linear.

Theorem 2. *There exists a Bayes covariant combining method for every $K \geq 3$.*

Proof. Consider parameterization of F given by $s_{ij} = \log(\frac{1}{r_{ij}} - 1)$ for $1 \leq i < j \leq K$. From (6) we have for the action of $\mathbf{q} = (q_1, q_2, \ldots, q_K)$:

$$s^{\mathbf{q}}_{ij} = s_{ij} + \log q_j - \log q_i \tag{25}$$

Thus in parameterization by s_{ij}, the group of reweighing vectors acts via translations. Note that since s_{ij} is a function of r_{ij} only, any symmetry properties of pairwise matrix $\mathbf{R}$ are preserved in passing to s_{ij} coordinates. Moreover, in s_{ij} coordinates:

- the set of feasible matrices is the $\binom{K}{2}$ dimensional real vector space,
- point P is just the origin of the vector space.

For $i \leq K$ consider the one-parameter subgroup $\mathbf{G}_i$ of $\mathbf{G}$ parameterized as:

$$\mathbf{q} = (\overbrace{1, 1, \ldots}^{i-1}, q, 1, 1, \ldots) \tag{26}$$

Let $h = (h_1, h_2)$ be a bijective mapping of the set $\{1, 2, \ldots, \binom{K}{2}\}$ onto the set of pairs $\{(i, j) \mid 1 \leq i < j \leq K\}$. Such a function induces ordering of coordinates s_{ij}, which will allow us to express tangent vectors to the Bradley–Terry manifold at P. Namely, for $m \leq \binom{K}{2}$ the mth component of the tangent vector $\mathbf{m}_k$ to P along the action of $\mathbf{G}_k$ is given by

$$(\mathbf{m}_k)_m = \begin{cases} -1 & \text{if } h_1(m) = k \\ 1 & \text{if } h_2(m) = k \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

Subgroups $\mathbf{G}_i$ generate $\mathbf{G}$ and therefore vectors $\mathbf{m}_k$ generate the tangent space of the Bradley–Terry manifold at P. However vectors $\mathbf{m}_k$ are not linearly independent, because the action of $\mathbf{q}$ and $c \cdot \mathbf{q}$ is the same. Omitting one of the vectors, say $\mathbf{m}_1$, from $\mathbf{m}_i$ we arrive at an explicit basis $\mathbf{M} = (\mathbf{m}_2, \mathbf{m}_3, \ldots, \mathbf{m}_K)$ of the tangent space of the Bradley–Terry manifold.

All that remains at this point is to solve the normal equations. Let $\mathbf{N}$ be a basis of the orthogonal complement to the tangent space, and let $\mathbf{s}$ represent a feasible matrix. We have:

$$\mathbf{s} = \mathbf{M}\mathbf{u} + \mathbf{N}\mathbf{v} \tag{28}$$

$$\mathbf{M}'\mathbf{s} = \mathbf{M}'\mathbf{M}\mathbf{u} + (\mathbf{M}'\mathbf{N})\mathbf{v} = \mathbf{M}'\mathbf{M}\mathbf{u} \tag{29}$$

$$(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{s} = \mathbf{u} \tag{30}$$

If we write $\mathbf{u}' = (u_1, \ldots, u_{K-1})$, then by (25), (24) and (28) the proposed Bayes covariant combining method is given by:

$$\hat{p}_i = \frac{\exp(u_{i-1})}{\sum_{i=1}^{K} \exp(u_{i-1})}, \quad \text{where } u_0 = 0. \tag{31}$$

Finally, note the special form of $\mathbf{M}'\mathbf{M}$. From (27) it follows that it has $(K - 1)$ on the diagonal and -1 elsewhere, which means it equals $K \cdot \mathbf{I} - \mathbf{j}\mathbf{j}'$, where $\mathbf{j} = (1, 1, \ldots)'$. Therefore:

$$(\mathbf{M}'\mathbf{M})^{-1} = \frac{1}{K}(\mathbf{I} + \mathbf{j}\mathbf{j}'), \tag{32}$$

allowing for simple computation of $\mathbf{u}$ via (30).

Please cite this article as: O. Šuch, S. Barreda, Bayes covariant multi-class classification, Pattern Recognition Letters (2016), http://dx.doi.org/10.1016/j.patrec.2016.08.014

6. Comparison to previously published methods

Let us compare the Bayes covariant method described above to previously suggested methods for combining pairwise classifiers. Two kinds of combining methods have been suggested, both of them derived from the Bradley–Terry model.

The first group of methods is based on minimization of a functional, for which the location of the optimum coincides with the solution of (3) when the latter exists. This underlies the HT method suggested by Hastie and Tibshirani [17], who propose to minimize the Kullback–Leibler divergence between r_{ij} and $\hat{p}_i/(\hat{p}_i+\hat{p}_j)$. Two further methods were suggested by Wu et al. [18], who propose to minimize the quadratic forms:

$$\min_{\mathbf{p}} \sum_{i=1}^{k} \left[\sum_{j:j\neq i}^{k} (r_{ij}p_j - r_{ji}p_i) \right]^2, \quad \text{for the WLW1 method} \tag{33}$$

$$\min_{\mathbf{p}} \sum_{i=1}^{k} \sum_{j:j\neq i}^{k} (r_{ij}p_j - r_{ji}p_i)^2 \quad \text{for the WLW2 method} \tag{34}$$

The second group of methods is characterized by attempting to solve (3) directly. The PKPD method of Price et al. [26] solves a consistent subset of (3) to arrive at $\hat{p}_i$ for each i, whereas the SBT method of Šuch et al. imposes a self-consistency condition on such estimates [27]. Let us abbreviate the Bayes covariant method described in Sections 3 and 4 as SB. All of these methods satisfy the 3-symmetry condition.

Theorem 3. *None of the HT, WLW1, WLW2, PKPD and SBT methods is Bayes covariant.*

Proof. We will use uniqueness of Bayes covariant classifier proved in Theorem 1. We can see from (15) that the ratio $A=\hat{p}_1/\hat{p}_2$ of the Bayes covariant classifier is not rational over the field generated by r_{ij}. Since all estimates of the methods except the HT method are rational over the field, it follows that the methods are not Bayes covariant.

To exclude the possibility that HT is Bayes covariant we numerically computed the Kullback–Leibler divergence for estimates given by both the HT and SB methods. When the matrix of pairwise likelihoods is [17, p. 452]:

$$\mathbf{R} = \begin{pmatrix} \cdot & 0.9 & 0.4 \\ 0.1 & \cdot & 0.7 \\ 0.6 & 0.3 & \cdot \end{pmatrix} \tag{35}$$

then the estimate of the posteriors made by HT method is $\hat{p}_{HT} \approx (0.481, 0.242, 0.277)$ and the Bayes covariant estimate is $\hat{p}_{SB} \approx (0.548, 0.192, 0.26)$. The Kullback–Leibler divergence is ≈0.376 for the former, whereas it is ≈ 0.397 for the latter. It follows from this that our method does not minimize the Kullback–Leibler divergence, and hence it is distinct from the HT method. □

To gain further insight into the relationship between the methods outlined here, we prepared multidimensional scaling (MDS) plots. For each $K \le 9$ we sampled 10^5 values for r_{ij} uniformly from the $\binom{K}{2}$-dimensional cube. Then we defined the distance between methods as the proportion of samples for which the MAP estimates of two methods differ. This distance is symmetric and respects the triangular inequality. We used classical MDS [28] to obtain the canonical two dimensional representation as shown in Fig. 1. As seen in the figure, the results for $K=3$ turned out quite differently from the others in that all the methods are arranged essentially on a single line. Moreover, the position of the PKPD and SBT methods were indistinguishable. In looking at the rest of the panels in Fig. 1 we can see that the Bayes covariant method is closest to the method of Hastie and Tibshirani in its classification.

Table 1
Predicted posteriors for pairwise comparison matrix **R** given by (36).

Combining method	Predicted posteriors $\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
WLW2	0.499863	0.499862	0.000275
SBT	0.499975	0.499750	0.000275
PKPD	0.500067	0.499842	0.000091
WLW1	0.500088	0.499638	0.000275
HT	0.500312	0.499413	0.000275
SB	0.682950	0.316902	0.000147

7. Practical considerations

Let us analyze the applicability of the Bayes covariant method in practice.

First and foremost, by construction it gives the correct solution of the Bradley–Terry equations whenever these equations have a solution. This is true of all other methods (WLW1, WLW2, HT, SBT, PKPD) and hence in this case all methods agree in predicted posteriors as well as in MAP classification. Second, the implementation of the Bayes covariant method described in Section 5 is straightforward and the resulting method is fast. Its computation requires computing $< K^2$ logarithms and exponentials and computation of two matrix–vector multiplications. The only methods which are simpler to implement are PKPD and the fast classification variant of HT method [17, Theorem 1]. Methods SBT and WLW1 require finding the eigenvector for eigenvalue $\lambda = 1$ of a Markov matrix, which can be done quickly. The algorithm for WLW2 method is based on a fast iterative scheme. The full HT method is the slowest.

A weaker point of the Bayes covariant method is numerical instability near the boundary of the subset of feasible matrices within the set of nonnegative matrices satisfying $\mathbf{A}+\mathbf{A}'=\mathbf{j}\mathbf{j}'-\mathbf{I}$. It may well happen in practice that a nondiagonal entry in matrix **R** approaches 0 or 1. In these cases the entry $s_{ij}=\log(\frac{1}{r_{ij}}-1)$ approaches infinity or negative infinity. This may cause instability in numerical computations, since a computer represents a real number with a finite mantissa and a finite exponent. A simple workaround is to eliminate from combining all classes C_i for which there exists $j\neq i$ such that r_{ij} is zero, or r_{ij} is very small, e.g. below 10^{-5}.

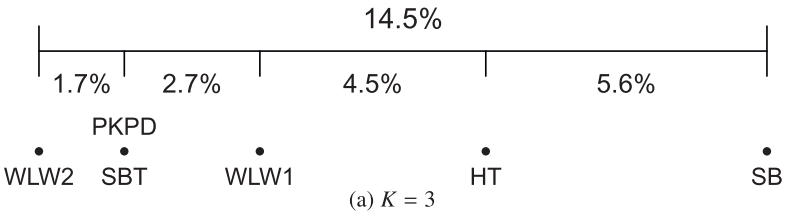
Another important consideration is the required precision of the binary predictions r_{ij}. Consider the following pairwise posterior matrix

$$\mathbf{R} = \begin{pmatrix} \cdot & 1/2 & 1-10^{-4} \\ 1/2 & \cdot & 1-10^{-3} \\ 10^{-4} & 10^{-3} & \cdot \end{pmatrix} \tag{36}$$

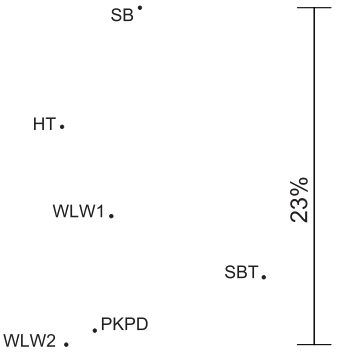
The third row contains very small entries, hence one may expect that class 3 is quite unlikely to be predicted (it lost binary comparisons with classes 1 and 2 by large margins). Furthermore, the binary decision between classes 1 and 2 ended up in a tie. As a result, one would guess that final posteriors would be approximately (1/2, 1/2, 0). The results of predictions by all considered methods are shown in Table 1.

All methods give preference to class 1, but only the Bayes covariant method does so by a large margin. The way to understand the preference for class 1 is as follows. On the one hand, pairwise comparison between classes 1 and 2 ended in a tie so one would expect that posteriors for classes 1 and 2 are approximately the

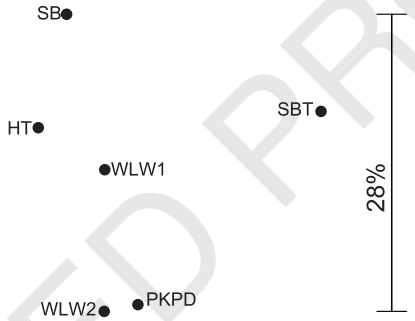
Please cite this article as: O. Šuch, S. Barreda, Bayes covariant multi-class classification, Pattern Recognition Letters (2016), http://dx.doi.org/10.1016/j.patrec.2016.08.014



(a) $K = 3$

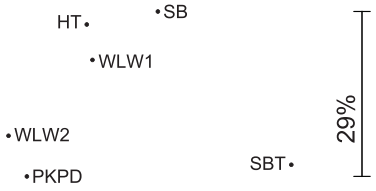


(b) $K = 4$

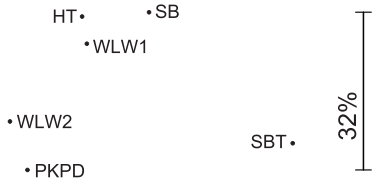


(c) $K = 5$

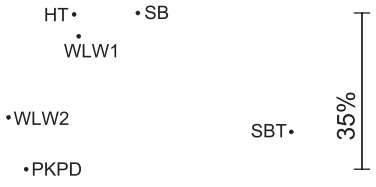
(d) $K = 6$



(e) $K = 7$



(f) $K = 8$



(g) $K = 9$

Fig. 1. Metric MDS visualizations of our method (SB) and the methods listed in Section 6. The underlying metric is the expected proportion of points in parameter space where a given pair of methods disagree in the prediction of the likeliest category.

former, the posterior is above 0.9 for the SB method, whereas it drops below 0.5 early on for other methods. In the latter case, the posterior is near 1.0 for the SB method whereas it does not exceed 0.8 for any of the other methods.
2. The overall posterior for /eh/ + /ae/ segments is similar to that of the other methods. This indicates that the method does not uniformly overshoot the posterior estimate. One can also observe that although $p(/ae/) + p(/eh)$ for the SB method closely mirrors the other methods, the $p(/eh/)$ component does not.
3. The posteriors oscillate less. Again, this can be seen both for /aa,ao/ and /iy/ segments. It is even more impressive given that the individual pairwise classifiers oscillate quite a bit as can be inferred from the posterior predictions from the different combining methods.
4. The duration of gaps where none of the three groups (/aa,ao/, /eh/ + /ae/ and /iy/) has posterior > 0.5 is much shorter for the SB method. This is mostly due to shortening of "uncertainty gap" between /eh/+/ae/ segments and /iy/ segments.

We can examine the last finding in more detail by analyzing the pairwise likelihood plots shown in Fig. 2c. We can see that the SB method gives a higher posterior ($\approx$0.5) at the onset of /iy/ than the pairwise LDA model between /eh/ and /iy/. The other methods do quite the opposite, they indicate lower posterior of /iy/ at the onset of /iy/ compared to the pairwise LDA model. This means that SB method extracted evidence for the presence of /iy/ from other pairwise models. One model, which detects earlier onset of /iy/ is the pairwise model classifying between /iy/ and /ae/. As we can see in Fig. 2c the model detects the onset of /iy/ earlier that /iy/-/eh/ model. Overall, extracting information from multiple comparisons r_{ij}, as demonstrated for the SB method, is a very desirable behavior for a combining method.

Methods

We considered 9 classes of monophthongs as described in the work of Lee and Hon [30]. For each class we extracted 10,000 samples from the train section of TIMIT sentences of male speakers. Sentences SA1 and SA2 were excluded. Then we trained

- a binary LDA model for each of the $\binom{9}{2} = 36$ pairs of classes,
- a multi-class LDA model,
- a multinomial model.

For training we used *MASS* and *nnet* R packages. The feature set consisted of 256 values in the log-periodogram. The log-periodogram was obtained by taking a 512-point window (at 16 kHz sampling rate its duration was 32 ms), which was weighted by Hanning window function, and then we took logarithms of magnitudes of individual Fourier coefficients.

9. Conclusion

We have introduced the notion of Bayes covariance for multi-class combing methods. We have shown that a Bayes covariant combining method exist for $K \geq 3$, and that for $K = 3$ there is a unique method given a 3-symmetry condition. We have compared the Bayes covariant method with five other combining methods. Somewhat surprisingly, these combining methods lack Bayes covariance, although they have the 3-symmetry property. The performance of the newly proposed method differs significantly from the other combining methods, both in Monte Carlo simulations as well as in a speech frame classification task. Crucially, the speech frame classification task demonstrated that the Bayes covariant method is able to extract information from all pairwise classifiers to arrive at more certain, less oscillatory posterior classification compared to previously suggested methods.

Acknowledgments

This w[illegible]was supported by Grants APVV-0219-12 and APVV-14-0560.

References

[1] L. Vincent, Google book search: Document understanding on a massive scale, in: Proceedings of the Ninth International Conference on Document Analysis and Recognition ICDAR 2007, vol. 2, 2007, pp. 819–823, doi:10.1109/ICDAR.2007.4377029.
[2] A. Graves, A.-R. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 6645–6649, doi:10.1109/ICASSP.2013.6638947.
[3] P.J. Lisboa, A.F. Taktak, The use of artificial neural networks in decision support in cancer: A systematic review, Neural Netw. 19 (4) (2006) 408–415, doi:10.1016/j.neunet.2005.10.007.
[4] W.E.H.D.J. Hand, Statistical classification methods in consumer credit scoring: A review, J. R. Stat. Soc. Ser. A (Stat. Soc.) 160 (3) (1997) 523–541. URL http://www.jstor.org/stable/2983268
[5] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297, doi:10.1007/BF00994018.
[6] V. Vapnik, The Nature of Statistical Learning Theory, Springer, 1996.
[7] R. Schapire, The boosting approach to machine learning: An overview, in: D. Denison, M. Hansen, C. Holmes, B. Mallick, B. Yu (Eds.), Nonlinear Estimation and Classification, Lecture Notes in Statistics, vol. 171, Springer, New York, 2003, pp. 149–171, doi:10.1007/978-0-387-21579-2_9.
[8] R. Rifkin, A. Klautau, In defense of one-vs-all classification, J. Mach. Learn. Res. 5 (2004) 101–141. URL http://dl.acm.org/citation.cfm?id=1005332.1005336
[9] O. Lzoray, H. Cardot, Comparing combination rules of pairwise neural networks classifiers, Neural Process. Lett. 27 (1) (2008) 43–56, doi:10.1007/s11063-007-9058-5.
[10] T. Dietterich, G. Bakiri, Solving multiclass learning problems via error-correcting output codes, J. Artif. Intell. Res. 2 (1995) 263–286, doi:10.1613/jair.105.
[11] E.L. Allwein, R.E. Schapire, Y. Singer, Reducing multiclass to binary: A unifying approach for margin classifiers, J. Mach. Learn. Res. 1 (2001) 113–141, doi:10.1162/15324430152733133.
[12] K. Crammer, Y. Singer, On the learnability and design of output codes for multiclass problems, Mach. Learn. 47 (2-3) (2002) 201–233, doi:10.1023/A:1013637720281.
[13] J. Fürnkranz, Round robin classification, J. Mach. Learn. Res. 2 (2002) 721–747, doi:10.1162/153244302320884605.
[14] C.-W. Hsu, C.-J. Lin, A comparison of methods for multiclass support vector machines, IEEE Trans. Neural Netw. 13 (2) (2002) 415–425, doi:10.1109/72.991427.
[15] Y. Shiraishi, K. Fukumizu, Statistical approaches to combining binary classifiers for multi-class classification, Neurocomputing 74 (5) (2011) 680–688, doi:10.1016/j.neucom.2010.09.004.
[16] D.J. Broad, H. Wakita, Piecewise-planar representation of vowel formant frequencies, J. Acoust. Soc. Am. 62 (6) (1977) 1467–1473, doi:10.1121/1.381676.
[17] T. Hastie, R. Tibshirani, Classification by pairwise coupling, Ann. Stat. 26 (2) (1998) 451–471, doi:10.1214/aos/1028144844.
[18] T.-F. Wu, C.-J. Lin, R.C. Weng, Probability estimates for multi-class classification by pairwise coupling, J. Mach. Learn. Res. 5 (2004) 975–1005. URL http://dl.acm.org/citation.cfm?id=1005332.1016791
[19] P. O'Connor, D. Neil, S.-C. Liu, T. Delbruck, M. Pfeiffer, Real-time classification and sensor fusion with a spiking deep belief network, Front. Neurosci. 7 (2013) 178, doi:10.3389/fnins.2013.00178.
[20] S. Hussain, S.-C. Liu, A. Basu, Improved margin multi-class classification using dendritic neurons with morphological learning, in: Proceedings of the 2014 IEEE International Symposium on Circuits and Systems (ISCAS), 2014, pp. 2640–2643, doi:10.1109/ISCAS.2014.6865715.
[21] S. Renals, S. King, Automatic speech recognition, in: W.J. Hardcastle, J. Laver, F.E. Gibbon (Eds.), Handbook of Phonetic Sciences, Wiley-Blackwell, 2013, pp. 804–838.
[22] T. Nearey, P. Assmann, M.-J. Solé, P.S. Beddor, M. Ohala, Probabilistic "sliding template" models for indirect vowel normalization, Experimental Approaches to Phonology, Oxford University Press, 2007, pp. 246–270.
[23] T.M. Nearey, P.F. Assmann, Modeling the role of inherent spectral change in vowel identification, J. Acoust. Soc. Am. 80 (5) (1986) 1297–1308, doi:10.1121/1.394433.
[24] J.M. Hillenbrand, T.M. Nearey, Identification of resynthesized /hvd/ utterances: Effects of formant contour, J. Acoust. Soc. Am. 105 (6) (1999) 3509–3523, doi:10.1121/1.424676.
[25] R.A. Bradley, M.E. Terry, Rank analysis of incomplete block designs: The method of paired comparisons, Biometrika 39 (3-4) (1952) 324–345, doi:10.1093/biomet/39.3-4.324.
[26] D. Price, S. Knerr, L. Personnaz, G. Dreyfus, G. Tesauro, D. Touretzky, T. Leen, Pairwise neural network classifiers with probabilistic outputs, Advances in Neural Information Processing Systems, vol. 7, MIT Press, 1995, pp. 1109–1116.

[27] O. Such, S. Benus, A. Tinajero, A new method to combine probability estimates from pairwise binary classifiers, Information Technologies Applications and Theory, ITAT, 2015, pp. 194–199. URL http://ceur-ws.org/Vol-1422/
[28] J.C. Gower, Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika 53 (3-4) (1966) 325–338, doi:10.1093/biomet/53.3-4.325.
[29] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, Darpa TIMIT acoustic phonetic continuous speech corpus CDROM, 1993, URL http://www.ldc.upenn.edu/Catalog/LDC93S1.html.
[30] K.F. Lee, H.W. Hon, Speaker-independent phone recognition using hidden Markov models, IEEE Trans. Acoust. Speech Signal Process 37 (11) (1989) 1641–1648, doi:10.1109/29.46546.

Please cite this article as: O. Šuch, S. Barreda, Bayes covariant multi-class classification, Pattern Recognition Letters (2016), http://dx.doi.org/10.1016/j.patrec.2016.08.014