The von Mises Naive Bayes Classifier for Angular Data

Pedro L. López-Cruz, Concha Bielza, and Pedro Larrañaga

Computational Intelligence Group
Departamento de Inteligencia Artificial
Facultad de Informática
Universidad Politécnica de Madrid
Campus de Montegancedo
Boadilla del Monte, 28660, Madrid, Spain
pedro.lcruz@upm.es, {mcbielza,pedro.larranaga}@fi.upm.es

Abstract. Directional and angular information are to be found in almost every field of science. Directional statistics provides the theoretical background and the techniques for processing such data, which cannot be properly managed by classical statistics. The von Mises distribution is the best known angular distribution. We extend the naive Bayes classifier to the case where directional predictive variables are modeled using von Mises distributions. We find the decision surfaces induced by the classifiers and illustrate their behavior with artificial examples. Two applications to real data are included to show the potential uses of these models. Comparisons with classical techniques yield promising results.

Keywords: Naive Bayes classifier, supervised classification, circular statistics, directional statistics, angular data, von Mises distribution.

1 Introduction

Scientists from a wide range of fields use angles to capture some properties of the phenomena they study, e.g., meteorologists analyze the direction of wind currents and waves, biologists measure the growth direction of plants and the movement of animals, etc.

Angular data have some distinctive properties that rule out the use of classical statistics. Therefore, common descriptive statistical tools have to be adapted to work with this kind of information, e.g., rose diagrams are used instead of regular histograms, the mean direction is computed taking into account the periodicity of the data, etc. Directional statistics [1,2] provides the theoretical background and the techniques to properly manage these data.

In this paper, we introduce the von Mises naive Bayes (vMNB) classifier for use with angular data. We review the naive Bayes classifier (NB) in Sect. 2, and the von Mises distribution in Sect. 3. Section 4 introduces vMNB, and its decision surfaces and properties are analyzed at length. Artificial examples are used to illustrate the behavior of the classifiers. Two applications to real data

J.A. Lozano, J.A. Gámez, and J.A. Moreno (Eds.): CAEPIA 2011, LNAI 7023, pp. 145–154, 2011.

and the statistical comparisons with classical techniques are included in Sect. 5. Finally, Sect. 6 concludes with a discussion and outlines future work. Detailed derivations of the formulas can be found in the Appendix.

2 The Naive Bayes Classifier

The NB classifier [3] is one of the best known models for supervised classification [4]. In NB, the class is modeled as a discrete variable C, and the set of its possible class values is noted $val(C)$. The set of predictive variables is $\{X_1, \ldots, X_n\}$. Figure 1 shows the graphical structure of the NB classifier, where the nodes represent the variables in the domain, and the arcs encode the conditional (in)dependence relationships between them [5]. NB assumes that the predictive variables are conditionally independent given the class value. NB uses a maximum a posteriori decision rule to classify the objects, i.e., it assigns each object to the class c^* with maximum posterior probability. Given an object with predictive variable values $\mathbf{x} = (x_1, \ldots, x_n)$, this is obtained as:

$$c^* = \arg \max_{c \in val(C)} p(C = c) \prod_{i=1}^{n} \rho(X_i = x_i | C = c),$$

where $\rho(\cdot)$ is a general probability function, i.e., a probability distribution $p(\cdot)$ for discrete variables or a probability density function $f_X(\cdot)$ for continuous variables.

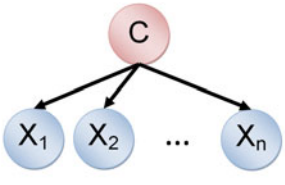


Fig. 1. Graphical structure of the naive Bayes classifier

Although conditional independence is a strong assumption, NB has been successfully applied to a wide range of problems [6], and its theoretical properties have been studied at length [7]. NB is a linear classifier when binary [3] or multinomial [8] predictive variables are used. On the other hand, the decision surfaces are polynomials when ordinal predictive variables are used [4].

3 The von Mises Distribution

The periodicity of angular data rules out the use of classical probability distributions. The most straightforward solution is to wrap linear distributions around the circle. Several distributions have been adapted according to this approach,

e.g., the wrapped normal distribution [9]. However, specific probability distributions have also been proposed for angular data. The von Mises distribution [10] is the best known circular distribution, as it is the circular analogue of the normal distribution. A variable Φ, defined in a circular domain $(-\pi, \pi]$, follows a von Mises distribution $vM(\mu_\Phi, \kappa_\Phi)$ if its probability density function is

$$f_\Phi(\phi; \mu_\Phi, \kappa_\Phi) = \frac{\exp(\kappa_\Phi \cos(\phi - \mu_\Phi))}{2\pi I_0(\kappa_\Phi)}, \tag{1}$$

where μ_Φ is the mean direction, $\kappa_\Phi \geq 0$ is the concentration of the points around the mean, and $I_\nu(\cdot)$ is the modified Bessel function of the first kind with order $\nu \in \mathbb{R}$, defined by

$$I_\nu(x) = \frac{1}{2\pi} \int_0^{2\pi} \cos(\nu\phi) \exp(x \cos\phi) d\phi \ .$$

The von Mises distribution is unimodal, with the mode (highest density) at μ_Φ and the antimode (lowest density) at $\mu_\Phi \pm \pi$. The distribution of the points around the circumference is uniform when $\kappa_\Phi = 0$, whereas high values of κ_Φ yield points tightly clustered around the mean. Given a sample of N points $\{\phi_1, \ldots, \phi_N\}$, the maximum likelihood estimators of the parameters in the distribution are the sample mean direction

$$\widehat{\mu}_\Phi = \arctan \frac{\overline{C}}{\overline{S}}, \text{ with } \overline{C} = \frac{1}{N} \sum_{i=1}^{N} \cos\phi_i, \text{ and } \overline{S} = \frac{1}{N} \sum_{i=1}^{N} \sin\phi_i, \tag{2}$$

and the sample concentration value

$$\widehat{\kappa}_\Phi = A^{-1}(\overline{R}), \text{ where } A(\widehat{\kappa}_\Phi) = \frac{I_1(\widehat{\kappa}_\Phi)}{I_0(\widehat{\kappa}_\Phi)} = \overline{R} = \sqrt{\overline{C}^2 + \overline{S}^2} \ . \tag{3}$$

Unfortunately, $\widehat{\kappa}_\Phi$ cannot be found analytically and approximations have to be computed numerically [2]. Figure 2 shows a sample of 100 points drawn from the distribution $\Phi \sim vM(\pi/4, 5)$ using the CircStat toolbox for MATLAB [11].

4 The von Mises Naive Bayes Classifier

In this section we introduce the vMNB classifier, where the conditional probability density functions of the predictive variables are modeled using von Mises distributions. The conditional probability densities for a variable Φ given the class value c are noted $(\Phi|C = c) \equiv \Phi^{(c)} \sim vM(\mu_\Phi^{(c)}, \kappa_\Phi^{(c)})$. We study the behavior of the classifier by deriving the decision surfaces it induces. We assume that the class is binary, e.g., $val(C) = \{1, 2\}$. When the class has more than two values, we have to compute the decision surface for each pair of values and label each subregion with the class having the maximum posterior probability. For detailed derivations of the decision surfaces included in this paper see the Appendix available at `http://cig.fi.upm.es/components/com_phocadownload/container/vmnbappendix.pdf`.

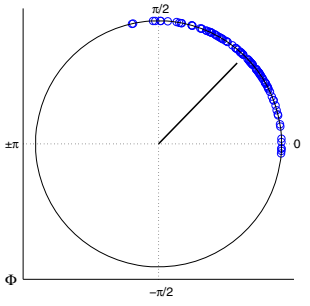


Fig. 2. Sample of 100 points drawn from the distribution $vM(\pi/4, 5)$. The black line represents the sample mean direction $\widehat{\mu}_\Phi$ and its length is the sample mean resultant length $\overline{R}$.

4.1 One Predictive Variable

We first analyze the simplest scenario where only one predictive variable Φ is used for classification. The decision surface induced by the classifier is computed by equaling the posterior probability distribution of the two class values

$$p(C = 1|\Phi = \phi) = p(C = 2|\Phi = \phi) \ . \tag{4}$$

By applying Bayes' rule and substituting the von Mises density (1) in (4), we get

$$\frac{p(C=1)}{2\pi I_0(\kappa_\Phi^{(1)})} \exp(\kappa_\Phi^{(1)} \cos(\phi - \mu_\Phi^{(1)})) = \frac{p(C=2)}{2\pi I_0(\kappa_\Phi^{(2)})} \exp(\kappa_\Phi^{(2)} \cos(\phi - \mu_\Phi^{(2)})) \ .$$

Simplifying, taking logarithms and operating, we finally get the two angles that bound the class subregions (see the Appendix):

$$\phi' = \alpha + \arccos(D/T)$$
$$\phi'' = \alpha - \arccos(D/T),$$

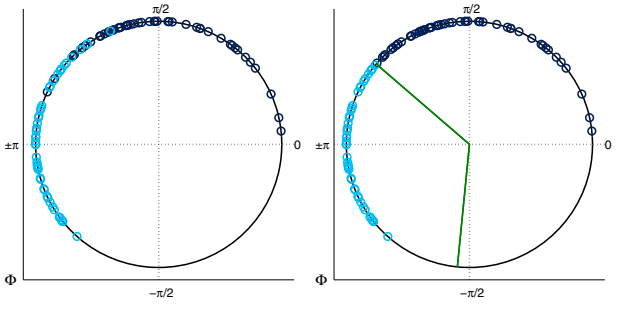
where $\cos\alpha = a/T$, $\sin\alpha = b/T$, $D = -\ln \frac{p(C=1)I_0(\kappa_\Phi^{(2)})}{p(C=2)I_0(\kappa_\Phi^{(1)})}$, $T = \sqrt{a^2+b^2}$, $a = \kappa_\Phi^{(1)} \cos\mu_\Phi^{(1)} - \kappa_\Phi^{(2)} \cos\mu_\Phi^{(2)}$, and $b = \kappa_\Phi^{(1)} \sin\mu_\Phi^{(1)} - \kappa_\Phi^{(2)} \sin\mu_\Phi^{(2)}$.

vMNB finds two angles that divide the circumference into two subregions, one for each class value. The two angles ϕ' and ϕ'' are defined with their bisector angle α, which depends on the mean directions $\mu_\Phi^{(1)}, \mu_\Phi^{(2)}$ and concentrations $\kappa_\Phi^{(1)}, \kappa_\Phi^{(2)}$, of Φ given each of the two class values. The distance between the angles also depends on both the concentration and the mean directions. Alternatively, we can substitute $(x, y) = (\cos\phi, \sin\phi)$ to compute the Cartesian coordinates of the decision surface that bounds the class subregions, obtaining the following expression (see the Appendix for details):

$$(\kappa_{\Phi}^{(1)}\mu_X^{(1)} - \kappa_{\Phi}^{(2)}\mu_X^{(2)})x - (\kappa_{\Phi}^{(1)}\mu_Y^{(1)} - \kappa_{\Phi}^{(2)}\mu_Y^{(2)})y - D = 0 \ . \tag{5}$$

Equation (5) defines a decision line that bounds the class regions. Therefore, vMNB with one predictive variable is a linear classifier.

We illustrate the behavior of the classifier with an artificial example. The class variable C is binary and its values are considered equiprobable a priori, i.e., $p(C = 1) = p(C = 2) = 0.5$. The conditional probability densities of Φ given each class value are $\Phi^{(1)} \sim vM(\pi/2, 2)$ and $\Phi^{(2)} \sim vM(\pi, 5)$. Figure 3(a) shows a sample of 100 points drawn from these distributions, whereas Fig. 3(b) shows the classification provided by vMNB and the decision angles that bound the class regions (green lines): $\phi' = 2.43$ (139.23°) and $\phi'' = -1.67$ (−95.63°).



(a) True classification of the points (b) Predicted classification using vMNB

Fig. 3. True class and class predicted using vMNB for a sample of 100 points. Points with $C = 1$ are shown in dark blue, whereas points with $C = 2$ are shaded light blue.

4.2 Two Predictive Variables

We can use the same approach to analyze the behavior of the classifier when two circular predictive variables Φ and Ψ are included in the model. In this scenario, the domain defined by the predictive variables is a torus $(-\pi, \pi] \times (-\pi, \pi]$. The decision surface induced by the vMNB classifier is given by

$$p(C = 1|\Phi = \phi, \Psi = \psi) = p(C = 2|\Phi = \phi, \Psi = \psi) \ . \tag{6}$$

By applying conditional independence, Bayes' rule, substituting the von Mises density function (1) in (6) and operating, we get

$$a\cos\phi + b\sin\phi + c\cos\psi + d\sin\psi + D = 0, \tag{7}$$

where a, b, c, d and D are constants (see the Appendix). The Cartesian coordinates of the points lying on the surface of a torus can be computed using:

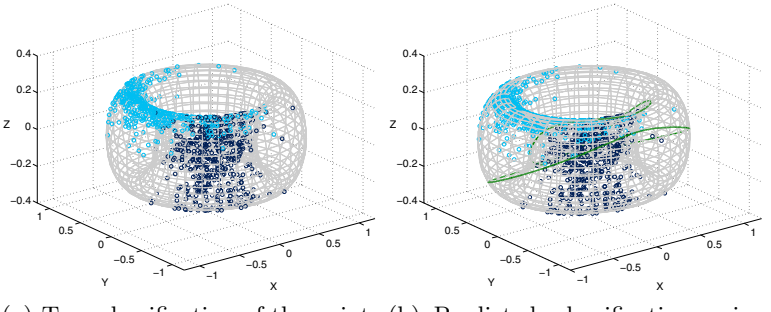
$$\begin{aligned} x &= (L + l\cos\phi)\cos\psi \\ y &= (L + l\cos\phi)\sin\psi \\ z &= l\sin\phi, \end{aligned} \tag{8}$$

where L is the distance from the center of the torus to the center of the revolving circumference that generates it, and l is the radius of the revolving circumference. Isolating the trigonometric functions in (8), replacing them in (7) and operating, we get the following decision surfaces:

$$clx + dly - az^2 + bz\sqrt{l^2 - z^2} + bLz + (aL + Dl)\sqrt{l^2 - z^2} + al^2 + DLl = 0,$$
$$clx + dly - az^2 - bz\sqrt{l^2 - z^2} + bLz - (aL + Dl)\sqrt{l^2 - z^2} + al^2 + DLl = 0 \ .$$

These decision surfaces are quadratic in z, so vMNB is a more complex and flexible classifier when two variables are included than when only one variable is used. This behavior is different in the NB with discrete predictive variables, where the complexity of the decision surfaces (hyperplanes) remains the same when the number of predictive variables is increased [8]. The decision surfaces are also hyperplanes when the predictive variables are statistically independent and modeled with Gaussian distributions that share the same variance. However, as far as we know, no result has been given in this particular scenario, where the predictive variables are conditionally independent given the class value and have different variances.

The following artificial example illustrates this behavior. Figure 4(a) shows a sample of 1000 points drawn using the distributions $\Phi^{(1)} \sim vM(\pi, 2)$ and $\Psi^{(1)} \sim vM(-2\pi/3, 6)$ for points in class $C = 1$, and distributions $\Phi^{(2)} \sim vM(\pi/2, 5)$ and $\Psi^{(2)} \sim vM(\pi, 3)$ for points in class $C = 2$. The classes are considered equiprobable a priori. The classification provided by vMNB and the complex decision boundaries that separate the two class regions are shown in Fig. 4(b).



(a) True classification of the points (b) Predicted classification using vMNB

Fig. 4. True class and class predicted using vMNB for a sample of 1000 points. Points with $C = 1$ are shown in dark blue, whereas points with $C = 2$ are shaded light blue. The decision boundaries are drawn in green.

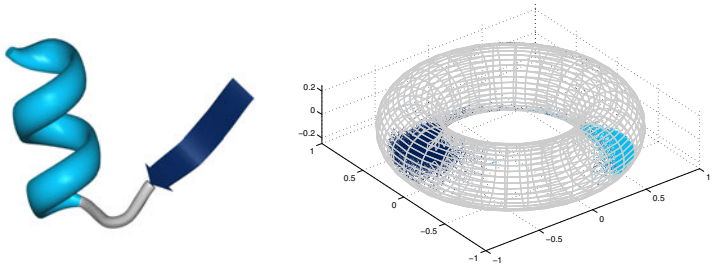
5 Experiments with Real Data

In this section, we apply the classifiers presented above to real world data from two different problems studied in biology:

Group Identification in Megaspores: In this problem we classify two groups of lycopsid megaspores based on the angle of orientation of the sporopollenin wall elements. The dataset is an example included in Oriana software (`http://www.kovcomp.co.uk/oriana`). It was first obtained and analyzed in [12]. The data are measured in degrees and represent the orientation of the element relative to a baseline drawn perpendicular to the spore surface. The two groups of megaspores used in this study are called *Selaginellalean* and *Isoetalean*. The dataset includes 960 entries, where 360 are *Selaginellalean* (37.5%) and 600 are *Isoetalean* (62.5%).

Protein Secondary Structure Prediction Using Dihedral Angles: The three dimensional structure of proteins is the key to identifying their function and behavior [13]. Many models tend to predict the protein secondary structure before modeling the tertiary structure. Dihedral angles (ϕ, ψ) are of key importance since they primarily define the protein's backbone conformation. In this example, we use the dihedral angle values of aminoacids to distinguish between the two most common secondary structures in proteins: the α-helix and the β-sheet. The data were retrieved from the Protein Geometry Database [13]. The dihedral angles for all the compositions corresponding to one residue were retrieved. We erased the instances with missing dihedral angles and selected the conformations corresponding to α-helices and β-sheets to obtain a dataset containing 49,676 instances. The number of instances for each class value were 28,141 α-helices (56.65%) and 21,535 β-sheets (43.35%). Figure 5(a) shows an α-helix (light blue) and a β-sheet (dark blue) conformation in a protein. Figure 5(b) shows the dihedral angles of all the aminoacids in α-helix (light blue) and β-sheet (dark blue) data conformations, mapped into a torus. Von Mises distributions have been used to model dihedral angles of protein structures in a number of works, e.g., [14,15].

We use vMNB to solve these problems. The maximum likelihood estimators of the parameters in Eq. (2) and (3) are computed using [11]. As far as we know, supervised classification problems using angular data as predictive information have not been systematically studied before. Therefore, we could not find any other approaches that manage directional data. We compare our results with the commonly used Gaussian NB classifier (GNB) and the multinomial NB classifier (mNB) using a supervised discretization algorithm [16]. The accuracy of the classifiers is estimated with a stratified 10-fold cross-validation procedure. Table 1 shows the classifiers' accuracies. We test if the difference in accuracy is significant by applying a right-tailed t-test over the sorted difference of accuracies in a 10-fold cross validation averaged over 10 runs, as recommended in [17]. Table 1 also shows the p-values of this t-test for each pair of classifiers (the first classifier is better than the second). In Megaspores dataset, we can only find statistical differences between GNB and mNB. On the other hand, vMNB



(a) Structure of an α-helix (left) and a β-sheet (right)

(b) Dihedral angles of α-helices (light blue) and β-sheets (dark blue) mapped into a torus

Fig. 5. Structure and dihedral angle distribution of α-helices and β-sheets structures

outperforms both GNB and mNB in Protein dataset. Figure 6 illustrates the difference between modeling protein dihedral angle ψ with a Gaussian or a von Mises conditional distribution for C=2. The Gaussian distribution ignores the periodicity of the data and yields different densities for angles 180° (0.24) and −180° (0.0), which refer to the same angle. Also, the von Mises distribution is more peaked. The log-likelihood for the von Mises distribution given the data is higher than for the Gaussian distribution (see the legend in Fig. 6).

Table 1. Mean accuracy and standard deviation of the classifiers computed with stratified 10-fold cross-validation (left). P-values of a right-tailed t-test to check whether the difference in accuracy is significant (right).

	vMNB	GNB	mNB	vMNB vs. GNB	vMNB vs. mNB	GNB vs. mNB
Megaspores	76.56 ± 4.26	76.46 ± 4.26	74.79 ± 5	0.7287	0.0607	0.0461
Protein	98.04 ± 0.18	97.64 ± 0.19	97.76 ± 0.24	0.0000	0.0001	0.9962

6 Discussion

In this paper, we introduced the vMNB classifier for use with angular and directional data. First, the NB classifier and the univariate von Mises distribution were reviewed. Then, we analyzed the behavior of vMNB when von Mises distributions are used to model the conditional probability distributions of the predictive variables. We derived the decision surfaces for one and two predictive variables and illustrated them with artificial examples. We showed that vMNB is a linear classifier when only one predictive variable is included. Also, we showed that the decision surfaces induced by vMNB are much more complex when two

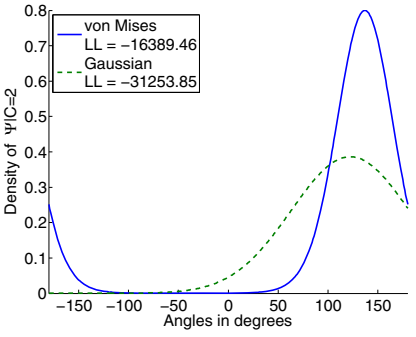


Fig. 6. Gaussian (dashed) and von Mises (solid) conditional density functions for ψ dihedral angles of class C=2 in Protein dataset

predictive variables are considered. Adding more predictive variables to vMNB can be easily done, and we could expect the complexity of the decision surfaces induced by these classifiers to grow accordingly. Two applications to real data from the field of biology were reported. The vMNB classifier achieved similar or better results than GNB and mNB in those datasets.

Conditional independence is a strong assumption, so a number of Bayesian classifiers that relax the NB assumption have been proposed, e.g., [18,19,20]. Extending vMNB to these Bayesian classifiers is not a straightforward matter. On the one hand, the conditional mutual information between variables modeled with von Mises distributions has to be computed in [19,20]. On the other hand, both marginal and conditional distributions of a multivariate von Mises cannot be von Mises distributions [21], making it difficult to model statistical dependencies between angular variables. Estimating the parameters of multivariate von Mises distributions is also challenging.

Hybrid scenarios combining discrete and continuous variables occur frequently in science. Classification models including categorical, Gaussian and von Mises distributions would account for a wide range of heterogeneous features, likely increasing the information available to the classifier and its accuracy. Learning and reasoning with these models is not trivial either.

We conclude that using von Mises distributions in Bayesian classifiers, and Bayesian networks generally, is both interesting and challenging. We hope that further research in this area will provide the tools necessary to properly manage directional data in machine learning.

Acknowledgments. This work has been supported by the Spanish Science and Innovation Ministry, Cajal Blue Brain Project (C080020-09), TIN2010-20900-C04-04 and Consolider Ingenio 2010-CSD2007-00018. PL L-C is supported by an FPU Fellowship (AP2009-1772) from the Spanish Education Ministry.

References

1. Fisher, N.I.: Statistical Analysis of Circular Data. Cambridge University Press (1993)
2. Mardia, K.V., Jupp, P.E.: Directional Statistics. John Wiley and Sons (2000)
3. Minsky, M.: Steps toward artificial intelligence. Proc. Inst. Radio. Eng. 49, 8–30 (1961)
4. Duda, R., Hart, P.: Pattern Classification and Scene Analysis. John Wiley and Sons (1973)
5. Koller, D., Friedman, N.: Probabilistic Graphical Models. Principles and Techniques. The MIT Press (2009)
6. Pourret, O., Naïm, P., Marcot, B.: Bayesian Networks: A Practical Guide to Applications. John Wiley and Sons (2008)
7. Domingos, P., Pazzani, M.: Beyond independence: Conditions for the optimality of the simple Bayesian classifier. Mach. Learn. 29, 103–130 (1997)
8. Peot, M.A.: Geometric implications of the naive Bayes assumption. In: Horvitz, E., Jensen, F.V. (eds.) Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence, pp. 414–419. Morgan Kauffman (1996)
9. Perrin, F.: Étude mathématique du mouvement Brownien de rotation. Ann. Sci. Ec. Norm. Super. 45, 1–51 (1928)
10. von Mises, R.: Über die "Ganzzahligkeit" der Atomgewichte und verwandte Fragen. Physikal. Z. 19, 490–500 (1918)
11. Berens, P.: CircStat: A MATLAB toolbox for circular statistics. J. Stat. Softw. 31(10), 1–21 (2009)
12. Kovach, W.L.: Quantitative methods for the study of lycopod megaspore ultrastructure. Rev. Palaeobot. Palynology 57(3-4), 233–246 (1989)
13. Berkholz, D.S., Krenesky, P.B., Davidson, J.R., Karplus, P.A.: Protein geometry database: A flexible engine to explore backbone conformations and their relationships to covalent geometry. Nucleic Acids Res. 38(suppl.1), D320–D325 (2010)
14. Mardia, K.V., Taylor, C.C., Subramaniam, G.K.: Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. Biometrics 63(2), 505–512 (2007)
15. Boomsma, W., Mardia, K.V., Taylor, C.C., Ferkinghoff-Borg, J., Krogh, A., Hamelryck, T.: A generative, probabilistic model of local protein structure. Proc. Natl. Acad. Sci. U.S.A. 105(26), 8932–8937 (2008)
16. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: Bajcsy, R. (ed.) Proceedings of the 13th International Joint Conference on Artificial Intelligence, pp. 1022–1027. Morgan Kaufmann (1993)
17. Bouckaert, R.R.: Estimating replicability of classifier learning experiments. In: Brodley, C.E. (ed.) Proceedings of the 21st International Conference on Machine Learning. ACM (2004)
18. Pazzani, M.J.: Searching for dependencies in Bayesian classifiers. Lecture Notes in Statistics 112, 239–248 (1995)
19. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Mach. Learn. 29, 131–163 (1997)
20. Sahami, M.: Learning limited dependence Bayesian classifiers. In: Simoudis, E., Han, J., Fayyad, U.M. (eds.) Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 335–338. AAAI Press (1996)
21. Mardia, K.V., El-Atoum, S.A.M.: Bayesian analysis for bivariate von Mises distributions. J. Appl. Stat. 37(3), 515–528 (2010)