AN ANALYSIS ON THE PERFORMANCE OF NAIVE BAYES PROBABILISTIC MODEL BASED CLASSIFIER FOR CARDIOTOCOGRAM DATA CLASSIFICATION

Sundar.C[1], M.Chitradevi[2] and G.Geetharamani[3]
[1]Christian College of Engineering and Technology, Oddanchatram, India.
Email ID: sundarc007@yahoo.com
[2]PRIST University Trichy Campus, Trichy, India. Email ID: chitra_win006@yahoo.co.in
[3]Anna University Chennai, BIT campus, Trichy, India.
Email ID: geeramdgl@rediffmail.com

ABSTRACT

Fetal heart rate (FHR) and uterine contractions (UC) are simultaneously recorded by Cardiotocography (CTG). The CTG, which is one of the most common diagnostic techniques used to evaluate maternal and fetal well-being during pregnancy and before delivery. By observing the Cardiotocography trace patterns doctors can understand the state of the fetus. There are several signal processing and computer programming based techniques for interpreting a typical Cardiotocography data. Even few decades after the introduction of cardiotocography into clinical practice, the predictive capacity of these methods remains controversial and still inaccurate. In this paper, we evaluate machine learning based Naine Bayes probabilistic model based classifier for their suitability towards classifying CTG data. We used Precision, Recall and F-Score as the metric to evaluate the performance. In previous works such as [1], the overall Precision, Recall and F-Score were only considered. But in this evaluation, we are going to measure class wise Precision, Recall and F-Score to make the analysis very specific. They arrived results prove that, even though the traditional clustering methods can identify the Normal CTG patterns, they were incapable of Suspicious and Pathologic patterns. This fact was not highlighted in [1].

KEYWORDS
Medical Data Classification, Cardiotocography, fetal heart rate, uterine contractions, Naive Bayes Probabilistic model.

1. INTRODUCTION

Data Mining (DM) and the technology of Knowledge Discovery from Data (KDD) has brought many new developments, methods, and technologies in the recent decade. Also the improvement of integration of techniques and the application of data mining techniques had contributed in handling of new kinds of data types and applications. However, the field of data mining and its application in medical domain is still young enough so that the possibilities of the application are still limitless [16].

The major challenges in medical domain are the extraction of intelligible knowledge from medical diagnosis data such as CTG data. Machine learning tools in medical diagnosis is gradually increased. The use of classification and recognition systems has improved with effectiveness to help medical experts in diagnosing diseases.

Cardiotocography (CTG)

Cardiotocography (CTG) is a technical means of recording the fetal heart rate (FHR) and the uterine contractions (UC) during pregnancy, typically in the third trimester to evaluate maternal

and fetal well-being. During the process of CTG analysis the FHR patterns are observed manually by obstetricians. In the recent past fetal heart rate baseline and its frequency analysis has been taken in to research on many aspects [2], [6].

Fetal heart rate (FHR) monitoring is mainly used to find out the amount of oxygen a fetus is acquiring during the time of labor [7]. Even then death and long term disablement occurs due to hypoxia during delivery. More than 50% of these deaths were caused by not recognizing the abnormal FHR pattern, even after recognizing not communicating the same without knowing the seriousness and the delay in taking appropriate action [7]. The currently proposed computation and datamining techniques for FHR can be used for analyzing and classifying the CTG data to avoid human mistakes and helps the doctors to take a decision. Computation and other datamining techniques can be used to analyze and classify the CTG data to avoid human mistakes and to assist doctors to take a decision.

Clustering and Classification

Data elements are placed into related groups without advance knowledge of the group definition is called as clustering. We use the very well used k-means algorithm that has been developed to efficiently solve the clustering problem. Normally clustering algorithm is used to form a group of objects whose positions are accurately known. The primary goal is to find an optimal method to divide the objects in to clusters [8]. Classification is a mining technique used to predict group membership for data instances.

2. PROBLEM DEFINITION

Cardiotocography (CTG), consisting of fetal heart rate (FHR) and tocographic (TOCO) measurements, is used to evaluate fetal well-being during the delivery. Since 1970, many researchers have worked different mining methods to help the doctors that interpret the CTG trace pattern from the field of signal processing and computer programming [2]. With the help of CTG trace pattern analysis the doctors with interpretations in order to reach a satisfactory level of reliability. So, they act as a decision support system in obstetrics. For everyday practice, none of them has been adapted worldwide. Baseline estimation in computer analysis of cardiotocographs , which is currently no consensus on the best methodology. More than 30 years after the introduction of antepartum cardiotocography into clinical practice, the predictive capacity of the method remains controversial. In a review of lot of articles published on this subject, it was found that its reported sensitivity varies between 2 and 100%, and its specificity between 37 and 100% [5]. So, in this work, we are going to evaluate two clustering algorithms for clustering CTG data.

The Medical Background of Cardiotocography (CTG)

Cardiotocography is a medical test conducted during pregnancy that records fetal heart rate (FHR) and uterine contractions. Either internal or external methods the tests may be conducted. During the internal testing, the uterus placed by a catheter after a specific amount of dilation has taken place. The external tests, a pair of sensory nodes are affixed to the mother's stomach. The CTG trace generally shows two lines. The fetal heart rate is recorded by the upper line in beats per minute and the uterine contractions are recording by the lower line from the TOCO.

Baseline Heart Rate

The baseline heart rate helps to evaluate the healthy functioning of the cardiovascular system. The baseline fetal heart rate is determined by approximating the mean FHR rounded to increments of 5 beats per minute (bpm) during a 10-minute window, excluding accelerations and decelerations and periods of marked FHR variability (greater than 25 bpm. Abnormal baseline is termed bradycardia and tachycardia.

The fluctuations are visually quantities as the amplitude of the peak- to-trough in bpm. Using this definition, the baseline FHR variability is categorized by the quantities amplitude as:

Absent- undetectable
Minimal- greater than undetectable, but less than or equal to 5 bpm
Moderate- 6 bpm - 25 bpm
Marked- greater than 25 bpm

Bradycardia: It is the resting heart rate of under 60 beats per minute, though it is seldom symptomatic until the rate drops below 50 beats/min. It may cause cardiac arrest in some patients
Tachycardia: It typically refers to a heart rate that exceeds the normal range for a resting heart rate (heart rate in an inactive or sleeping individual). Depending on the speed and type of rhythm, it can be dangerous.

Type 1 (early)
This occurs during the peak of the uterine contraction. The FHR with onset early in the contraction and return to baseline at the end of the contraction will be uniform, repetitive and periodic slowing. The reasons behind this may be fetal head compression, cord compression or early hypoxia. This occurs in first and second stage labor with decent of the head [4]. This is synchronous with uterine contraction.

Type 2 (late)
This occurs after the peak of the uterine contraction. The FHR with onset mid to end of the contraction and nadir more than 20 seconds after the peak of the contraction and ending after the contraction will also be uniform, repetitive and slowing. If the lag time is high seriousness is also high. This is also synchronous with uterine contraction. Mx: a fetal pH measurement is mandatory [4].

Type 3 (variable)
This is variable, repetitive, periodic slowing of FHR with rapid onset and recovery. Variable and isolated time relationships with contraction cycles may occur. Deceleration patterns in timing and shape resembles other types in some cases. If they occur consistently, there is a chance of fetal hypoxia. This is unrelated to uterine contractions. Mx: check fetal pH if the pattern persists after turning the patient on her side (or if other adverse features are present) [4].

3. THE NAIVE BAYES PROBABILISTIC MODEL

The probability model for a classifier is a conditional model [18],[19].

$$p(C|F_n, \dots, F_n)$$

Over a dependent class variable C with a small number of outcomes or classes, conditional on several feature variables F1 through Fn. The problem is that if the number of features n is large or when a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable.
Using Baye's theorem, we write

$$p(C|F_1,..,Fn) - \frac{p(C)p(F1,...,Fn|C)}{p(F1,..,Fn)}$$

In plain English the above equation can be written as

$$\text{Posterror} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

In practice we are only interested in the numerator of that fraction, since the denominator does not depend on C and the values of the features Fi are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model

$$p(C, F_n, \dots, F_n)$$

Which can be rewritten as follows, using the chain rule for repeated applications of the definition of conditional probability:

$$\begin{aligned}
&p(C, F_n, \dots, F_n) \\
&\propto p(c)p(F_1, \dots, F_n|C) \\
&\propto p(C)p(F_1|C)p(F_2, \dots, F_n|C, F_1) \\
&\propto p(C)p(F_1|C)p(F_2|C, F_1)p(F_3, \dots, F_n|C, F_1, F_2) \\
&\propto p(C)p(F_1|C)\, p(F_2|C, F_1)\, p(F_3|C, F_1, F_2)\, p(F_4, \dots, F_n|C, F_1, F_2, F_3) \\
&\propto p(C)p(F_1|C)\, p(F_2|C, F_1)\, p(F_3|C, F_1, F_2) \dots p(F_n|C, F_1, F_2, F_3, \dots, F_{n-1})
\end{aligned}$$

Now the "naive" conditional independence assumptions come into play: assume that each feature Fi is conditionally independent of every other feature Fj for i ≠ j given the class C. This means that

$$p(F_i|C, F_j) = p(F_i|C)$$

For i ≠ j , and so the joint model can be expressed as

$$\begin{aligned}
p(C|F_1, \dots, F_n) &\propto p(C)p(F_1|C)p(F_2|C)p(F_3|C) \dots \\
&\propto p(C)\prod_{i=1}^{n} p(F_i|C)
\end{aligned}$$

This means that under the above independence assumptions, the conditional distribution over the class variable C can be expressed like this:

$$p(C|F_1, \dots, F_n) = \frac{1}{Z}p(C)\prod_{i=1}^{n} p(F_i|C)$$

Where Z (the evidence) is a scaling factor dependent only on F1, …, Fn, i.e., a constant if the values of the feature variables are known.

Models of this form are much more manageable, since they factor into a so-called class prior p(C) and independent probability distributions p(Fi | C). If there are k classes and if a model for each p (Fi|C=c) can be expressed in terms of r parameters, then the corresponding naive Bayes model has (k − 1) + n r k parameters. In practice, often k = 2 (binary classification) and r = 1(Bernoulli variables as features) are common, and so the total number of parameters of the naive Bayes model is 2n+1, where n is the number of binary features used for classification and prediction.

4. CONSTRUCTING A CLASSIFIER FROM THE PROBABILITY MODEL

The discussion so far has derived the independent feature model, that is, the naive Bayes probability model [19],[20]. The naive Bayes classifier combines this model with a decision rule.

One common rule is to pick the hypothesis that is most probable; this is known as the maximum a posteriori or MAP decision rule. The corresponding classifier is the function classify defined as follows:

$$classify\ (f_1, \dots, f_n) = argmax\ p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c)$$

The Metrics Used for the Evaluation

Precision, recall and F-Score are computed for every (class, cluster) pair. But Rand index is a metric which will consider all the classes and the clusters as the whole.

Rand Index

The Rand index or Rand measure is a commonly used technique for measure of such similarity between two data clusters.

Given a set of n objects S = {O1, ..., On} and two data clusters of S which we want to compare: X = {x1, ..., xR} and Y = {y1, ..., yS} where the different partitions of X and Y are disjoint and their union is equal to S; we can compute the following values:

a is the number of elements in S that are in the same partition in X and in the same partition in Y,
b is the number of elements in S that are not in the same partition in X and not in the same partition in Y,
c is the number of elements in S that are in the same partition in X and not in the same partition in Y,
d is the number of elements in S that are not in the same partition in X and in the same partition in Y.

Intuitively, one can think of a + b as the number of agreements between X and Y and c + d the number of disagreements between X and Y. The Rand index, R, then becomes,

$$RI = \frac{a+d}{a+b+c+d}$$

The Rand index has a value between 0 and 1 with 0 indicating that the two set of data clusters do not agree on any pair of points and 1 indicating that the two data clusters are exactly similar.

Precision

Precision is calculated as the fraction of correct objects among those that the algorithm believes belonging to the relevant class. It can be loosely equated to accuracy and it will roughly answer the question: "How many of the points in this cluster belong there/ correctly classified?"
The Precision is calculated as:

$P(L_r, S_i) = n_{ri}/n_i$
for
class L_r of size n_r
cluster S_i if size n_i
n_{ri} data points in S_i from class L_r

Recall

Recall roughly answers the question: "Did all of the documents that belong in this cluster make it in?" In other words, recall is the fraction of actual objects that were identified.

The recall is calculated as:

$$R(L_r, S_i) = n_{ri}/n_r$$

F-Score

F-Score is the harmonic mean of Precision and Recall and will tries to give a good combination of the two. It is calculated with the equation:

$$F(L_r, S_i) = \frac{2 * R(L_r, S_i) * P(L_r, S_i)}{R(L_r, S_i) + P(L_r, S_i)}$$

Validating the Performance of the Classification

Classifier performance depends on the characteristics of the data to be classified. Performance of the selected algorithms is measured for Rand Index, Precision, Recall and F-Measure. Various empirical tests can be performed to compare the classifier like holdout, random sub-sampling, k-fold cross validation and bootstrap method. Here we did Holdout Cross validation for evaluating the proposed classification models.

Holdout Cross validation (It is equal to k-Fold Validation with k=2)

The holdout method is the simplest kind of cross validation. This 2-fold cross validation is the simplest variation of k-fold cross-validation. For each fold, we randomly assign data points to two sets d0 and d1; so that both sets are equal size (this is usually implemented by shuffling the data array and then splitting it in two). We then train on d0 and test on d1, followed by training on d1 and testing on d0. The advantage of this method is that it is usually preferable to the residual method and takes no longer to compute. However, its evaluation can have a high variance. The evaluation may depend heavily on which data points end up in the training set and which end up in the test set, and thus the evaluation may be significantly different depending on how the division is made .

This has the advantage that our training and test sets are both large, and each data point is used for both training and validation on each fold.

We used Holdout Cross validation (or k-Fold Validation with k=2) because, the dataset contains sufficient amount of samples which can be separated and used for training and testing (50%, 50%).

Further, instead of doing holdout cross validation for one time, the data set is randomly permuted and the training and testing records were randomly taken for 10 times and the average result of 10 such holdout cross validations were only considered.

Implementation and Evaluation

For implementing and evaluating the Naïve Bayes probabilistic model based classifier using Matlab 7.

5. RESULTS AND DISCUSSION

Data Set Information

For evaluating the algorithms under consideration, we used Cardiotocogram data from UCI Machine Learning Repository.

This data set contains 2126 fetal Cardiotocogram belonging to different classes. The data contains 21 attributes and two class labels. The CTGs were classified by three expert obstetricians and a consensus classification label assigned to each of them. Classification was both with respect to a morphologic pattern (A, B, C. ...) and to a fetal state (N, S, P). Therefore the dataset can be used either for 10-class or 3-class experiments. Here we use this data set for these evaluations.

Attribute Information

1. LB - FHR baseline (beats per minute)
2. AC - # of accelerations per second
3. FM - # of fetal movements per second
4. UC - # of uterine contractions per second
5. DL - # of light decelerations per second
6. DS - # of severe decelerations per second
7. DP - # of prolongued decelerations per second
8. ASTV - percentage of time with abnormal short term variability
9. MSTV - mean value of short term variability
10. ALTV - percentage of time with abnormal long term variability
11. MLTV - mean value of long term variability
12. Width - width of FHR histogram
13. Min - minimum of FHR histogram
14. Max - Maximum of FHR histogram
15. Nmax - # of histogram peaks
16. Nzeros - # of histogram zeros
17. Mode - histogram mode
18. Mean - histogram mean
19. Median - histogram median
20. Variance - histogram variance
21. Tendency - histogram tendency
22. CLASS - FHR pattern class code (1 to 10)
23. NSP - fetal state class code (Normal=1; Suspect=2; Pathologic=3)

Class Information

We used the data for a three class classification problem. The descriptions for the three classes are:

Normal
A CTG where all four features fall into the reassuring category.

Suspicious
A CTG whose features fall into one of the non-reassuring categories and the reassuring category and the remainder of features are reassuring.

Pathological
A CTG whose features fall into two or more of the Non-reassuring the reassuring category or two or more abnormal categories.

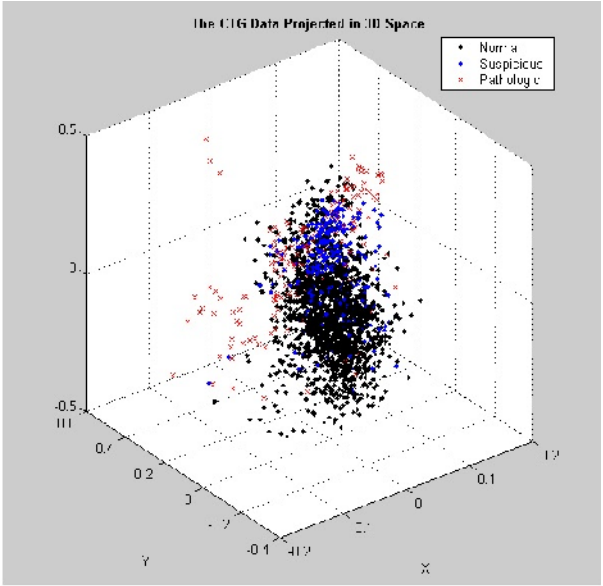

Figure 1: The 3D projection of CTG data

The Visualization of Data Space

The Figure1 image shows the projection of this 21 attribute (dimension) data in to a virtual three dimensional data space. We used three principal components of the data for this projection. In this plot, the normal CTG data points are shown in black dots, the suspicious data points are shown as blue dots and the Pathologic data points are shown as red 'x' mark. This figure roughly shows the distribution of the data in the virtual space.

The Numerical Results

The following table shows the average performance of Naive Bayes Probabilistic model based clustering algorithm. We tabulate the average results of ten trials.

Table 1: The Average Performance of Naive Bayes Probabilistic model based classifier

Metric	Normal	Suspicious	Pathological
Precision	0.975	0.473	0.574
Recall	0.833	0.831	0.636
F-Score	0.898	0.603	0.604

The Analysis of Results

The following chart obviously shows the performance of Naïve Bayes probabilistic model based classifier. It gives good precision for normal records and average performance in all other cases.

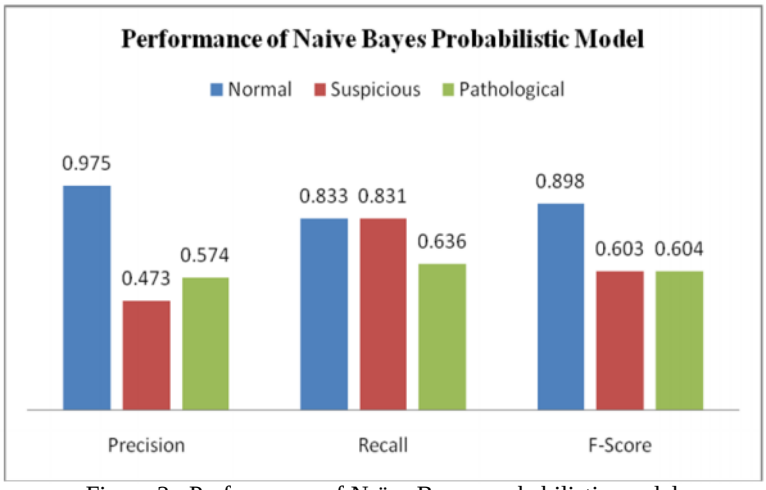

Figure 2: Performance of Naïve Bayes probabilistic model

The derived results clearly show Naïve Bayes probabilistic model based classifier can be used for the classification of CTG data. We realized that there are some training glitches in the case of suspicious records which caused.

6. CONCLUSION

We have evaluated the performance of Naïve Bayes probabilistic model based classification method with respect to Precision, Recall and F-Score. In previous works such as [1], the overall Precision, Recall and F-Score were only considered. But in this evaluation, we considered class-wise Precision, Recall and F-Score to make the analysis very specific. If we consider only the precision as a metric, then arrived results proves that, even though the machine learning based methods can distinguish the Normal CTG patterns from the Suspicious and Pathologic patterns with respect to precision and pathologic, but, they were incapable of distinguishing Suspicious. This fact was not highlighted in [1].

That is why we are getting comparatively poor average performance while classifying suspicious records with respect to precision. It is a major weakness of the algorithms which should be overcomes in future design. One may address the way to improve the system for getting proper results with different classes of CTG patterns. One may consider machine learning based method to design the CTG data classification system. Future works may address hybrid models using statistical and machine learning techniques for improved classification ac curacy.

REFERENCES

[1] Xiaojun Chen, Yunming Ye, Xiaofei Xu, Joshua Zhexue Huang ,2012, “A feature group weighting method for subspace clustering of high-dimensional data”, Pattern Recognition Vol.45,434-446, Elsevier

[2] Shahad Nidhal, M. A. Mohd. Ali1 and Hind Najah, 2010,“A novel cardiotocography fetal heart rate baseline estimation algorithm”, Scientific Research and Essays Vol. 5(24), p. 4002-4010.

[3] ANA. KLIMEŠOVÁ, EVA OCELÍKOVÁ, Multidimensional Data Classification, Proceedings of the 10th WSEAS International Conference on AUTOMATION & INFORMATION, ISSN: 1790-5117, ISBN: 978-960-474-064-2

[4] Stirrat, Mills and Draycott,2003, "Notes on Obstetrics and Gynaecology for the MRCOG, 5th Edition", ISBN: 9780443072239

[5] Diogo Ayres-de-Camposa, Cristina Costa-Santosb, Joa˜o Bernardesa, 2005, "Prediction of neonatal state by computer analysis of fetal heart rate tracings: the antepartum arm of the SisPorto1 multicentre validation study", European Journal of Obstetrics & Gynecology and Reproductive Biology 118, 52-60.

[6] http://www.academicjournals.org/SRE, ISSN 1992 – 2248 © 2010 Academic Journals.

[7] Antonia Costa, MD; Diogo Ayres-de-Campos, PhD; Fernada Costa, MD; Cristina Santos, MS; Joao Bernardes, PhD, 2009,"Prediction of neonatal academia by Computer analysis of fetal heart rate and ST event sibnals" AJOG – American Journal of Obstetrics and Gynecology.

[8] Ben Kao, Sau Dan Lee, Foris K.F.Lee, David W. Cheung, Wai-Shing Ho,2010," Clustering Uncertain Data using Voronoi Diagrams and R-Tree Index" IEEE Transactions on Knowledge and Data Engineering, Vol. 22(9), p. 1219 – 1233.

[9] E. Ocelikova, D. Klimesova, " Bays Classifier in multidimensional data classification " 15th Int. Conference Process Control 2005, pp. 188-1 – 188-5. Strbske Pleso, Slovakia.

[10] E. Ocelikovć, J Krištof, "Classification of multispectral data" Zbornik radova, Volume 25, Number 1(2001).

[11] Frank, A. Asuncion, UCI Machine Learning Repository. {http://archive. ics.uci.edu/ml}, 2010.

[12] Zhaohong Deng , Kup-Sze Choi , Fu-Lai Chung , Shitong Wang, 2010,"Enhanced soft subspace clustering integrating within-cluster and between-cluster information, Pattern Recognition", v.43 n.3, p.767-781, [doi>10.1016/ j.patcog. 2009.09.010]

[13] Hans-Peter Kriegel , Peer Kröger , Arthur Zimek, Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering, ACM Transactions on Knowledge Discovery from Data (TKDD), v.3 n.1, p.1-58, March 2009 [doi>10.1145/1497577.1497578]

[14] http://www.mathworks.in/help/toolbox/nnet/ug/bss33y1-1.html.

[15] S.Angle Latha Mary, K.R.Shankar Kumar," Evaluation of Clustering Algorithm with Cluster Validation Metrics" European Journal of Scientific Research ISSN 1450-216X Vol.69 No.1 (2012), p.61-72

[16] YI PENG, GANG KOU"A descriptive framework for the field of data mining and knowledge discovery" International Journal of Information Technology & Decision Making Vol. 7, No. 4 (2008) pp. 639–682.

[17] Michael Lloyd-Williams, 1997,"Discovering the hidden secrets in your data - the data mining approach to information", Information Research, {http:// informationr.net/ir/3-2/paper36.html}, Vol. 3 No.2.

[18] Caruana, R.; Niculescu-Mizil, A. (2006). "An empirical comparison of supervised learning algorithms". Proceedings of the 23rd international conference on Machine learning. CiteSeerX: 10.1.1.122.5901.

[19] George H. John and Pat Langley (1995). Estimating Continuous Distributions in Bayesian Classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. pp. 338-345. Morgan Kaufmann, San Mateo.

[20] Webb, G. I., J. Boughton, and Z. Wang (2005). Not So Naive Bayes: Aggregating One-Dependence Estimators. Machine Learning 58(1). Netherlands: Springer, pages 5–24.