Improving the Accuracy of Multiple Disease Detection System Using Novel Hybrid Classifier

[1] **Pranita Paradkar,** [2] **Sapna Choudhary**
[1] Student, M.Tech, [2] Professor, HOD
[1, 2]Department of Computer Science and Engineering,
[1, 2] Shri Ram Group of Institution (SRGI), Jabalpur, Madhya Pradesh (India)

Abstract: Automated disease detection from human body parameters has been a topic of medical research for more than a decade now. Many researchers have proposed various techniques for performing this task, but most of the proposed approaches work under certain disease conditions, and there are very few techniques which produce accurate results for multiple diseases. In the proposed approach, we are detecting 4 diseases and using a hybrid of support vector machine (SVM), k-Nearest Neighbour (kNN) and Naive Bayes classifiers in order to evaluate the diseases from input human body parameters. The evaluation results show more than 15% improvement in classification accuracy when compared to kNN and Naive Bayes classifiers individually. The proposed system can be extended for any number of diseases.

Keywords: Disease Detection, kNN, SVM, Naive Bayes, Hybrid, Multiple Diseases

1. INTRODUCTION

Automated disease detection from user's body parameters has been a topic of study for many researchers. The concept behind this automated detection is divided into the following steps [1],

- Dataset collection
- Parameter selection
- System training
- System evaluation

The dataset collection involves, collection of various datasets related to medical disease detection, these datasets can be collected from various online repositories like UCI machine learning repository, Stanford medical learning datasets and many others. These datasets contain variety of disease sets, which have the information about the human body parameters for a given disease and their inference in terms of the type of disease the body is infected with based on the given input parameters. Datasets ranging from simple diseases like fever to complex diseases like cancer and heart based arrhythmia diseases are available on these repositories.

The datasets provide various parameters for disease detection [2], these parameters are mainly human body parameters which include age, gender, blood pressure, heart rate and many more. Selection of most varying parameters from this set is done, in order to select the parameters which would change based on the disease which the person is suffering from, and thus would help to build a better classifier.

The system training part needs a good classifier [3], the training of the classifier is of utmost importance. The input dataset is divided into training and testing parts, where the training part is usually more than 50% of the dataset entries, while the remaining part is the testing part. A well trained classifier is the one which can correctly classify all or most of the testing set entries. The better the classifier classifies the entries correctly, the more is the accuracy of the classifier. A classifier with more than 90% accuracy is termed as a good classifier for disease detection in clinical research.

Knowledge discovery in databases is well-defined method consisting of many distinct steps. Data processing is that the core step, which ends within the discovery of hidden however helpful information from large databases. A proper definition of data} discovery in databases is given as follows: “Data mining is that the non trivial extraction of implicit antecedently unknown and doubtless helpful information concerning data” [4]. Data processing technology provides a user-oriented approach to novel and hidden patterns within the knowledge. The discovered information will be utilized by the care directors to boost the standard of service.

The discovered information may be utilized by the medical practitioners to cut back the amount of adverse drug result, to recommend more cost-effective the rapeutically equivalent alternatives. Anticipating patient’s future behavior on the given history is one amongst the vital applications of information mining techniques which will be employed in health care management. A major challenge facing care organizations (hospitals, medical centers) is that the provision of quality services at cheap prices. Quality service implies designation patients properly and administering treatments that are effective. Unhealthy clinical choices will result in black consequences that are thus unacceptable. Hospitals should conjointly minimize the price of clinical tests. They’ll succeed these results by using acceptable computer-based info and/or call support systems. Health care knowledge is huge. It includes patient central knowledge, resource management knowledge and reworked knowledge. Health care organizations should have ability to research knowledge.

Treatment records of countless patients will be keeping and processed and data processing techniques could facilitate in respondent many vital and significant queries involving health care. The provision of integrated info via the large patient repositories, there's a shift within the perception of clinicians, patients and payers from qualitative image of clinical knowledge by stern a lot of quantitative assessment of information with the supporting of all clinical and imaging data. As an example it would currently be doable for the physicians to match diagnostic info of assorted patients with identical conditions. Likewise, physicians may make sure their findings with the conformity of different physicians coping with a consistent case from everywhere the planet [5].

Diagnosis is taken into account as a big none the less involved task that must be distributed exactly and with efficiency. The automation of an equivalent would be extremely useful. Clinical choices are typically created supported doctors' intuition and skill instead of on the knowledge made knowledge hidden within the info. This applies results in unwanted biases, errors and excessive medical prices that affect the standard of service provided to patients. Wu, et al projected that integration of clinical call support with computer based patient records might scale back medical errors, enhance patient safety, decrease unwanted apply variation, and improve patient outcome [6]. This suggestion is promising as knowledge modeling and analysis tools, e.g., data processing, have the potential to come up with a knowledge-rich atmosphere which may facilitate to considerably improve the standard of clinical choices.

2. LITERATURE REVIEW

There is a unit varied data processing techniques offered with their suitableness obsessed on the domain application. Statistics offer a powerful basic background for quantification and analysis of results. However, algorithms supported statistics have to be compelled to be changed and scaled before they're applied to data processing. We have a tendency to currently describe some Classification data processing techniques with illustrations of their applications to health care. In rule set classifiers, complicated call trees may be tough to grasp, parenthetically as a result of data concerning one category is sometimes distributed throughout the tree. C4.5 introduced an alternate formalism consisting of an inventory of rules of the shape "if A and B and C and ...then category X", wherever rules for every category area unit sorted along. A case is classed by finding the primary rule whose conditions area unit glad by the case; if no rule is glad, the case is appointed to a default category. Decision tree embody CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and C4.5. These algorithms dissent in choice of splits, once to prevent a node from cacophonic, and assignment of sophistication to a non-split node [7]. CART uses Gini index to live the impurity of a partition or set of coaching tuples [6]. It will handle high dimensional categorical knowledge. call Trees may also handle continuous knowledge (as in regression) however they need to be regenerate to categorical knowledge. we'll discuss with a row as an information instance. the info set contains 3 predictor attributes, namely Age, Gender, Intensity of symptoms and one goal attribute, specifically illness whose values (to be foreseen from symptoms) indicates whether or not the corresponding patient have a particular illness or not.

In order to line the transformation parameters we have a tendency to should discuss attributes admire heart vessels. The LAD, RCA, LCX and lumen numbers represent the share of vessel narrowing (or blockage) compared to a healthy artery. Attributes LAD, LCX and RCA were divided by cut off points at fifty and seventieth. Within the medical specialty field, a seventieth price or higher indicates vital coronary illness and a five hundredth price indicates borderline illness. a worth less than five hundredth suggests that the patient is healthy. The foremost common cut off price employed by the medical specialty community to tell apart healthy from sick patients is five hundredth. The lumen artery is treated totally different as a result of it poses higher risk than the opposite 3 arteries. Attribute lumen was divided at thirty and five hundredth. the explanation behind these numbers is each the LAD and therefore the LCX arteries branch from the lumen artery then a defect in lumen is additional seemingly to cause a bigger morbid heart region. That is, narrowing (blockage) within the lumen artery is probably going to supply additional illness than blockages on the opposite arteries. That's why its cut off values area unit set 2 hundredth less than the opposite vessels. The 9 heart regions (AL, IL, IS, AS, SI, SA, LI, LA, AP) were divided into 2 ranges at a cut off purpose of zero.2, that means a insertion measure larger or equal than zero. 2 indicated a severe defect. CHOL was divided with cut off points two hundred (warning) and 250 (high). These values correspond to legendary medical settings. The design of the neural network utilized in this study is that the multi-layered feed-forward specification with twenty input nodes, ten hidden nodes, and ten output nodes. The quantity of input nodes is set by the finalized data; the quantity of hidden nodes is set through trial and error and therefore the number of output nodes is delineated as a spread showing the illness classification. The foremost wide used neural-network learning technique is that the BP algorithmic rule [8]. Learning during a neural network involves modifying the weights and biases of the network so as to attenuate a value operate. The value operate invariably includes a blunder term a live of however shut the network's predictions area unit to the category labels for the examples within the coaching set. In addition, it's going to embody a complexness term that reacts to a previous distribution over the values that the parameters will take. Neural networks are projected as helpful tools in higher cognitive process during a form of medical applications. Neural networks can ne'er replace human specialists however they will facilitate in screening and might be employed by specialists to see to it their designation. In general, results of illness classification or prediction task area unit true solely with a particular likelihood.

Another technique specifically neuro fuzzy is employed wide for classification; during this technique random back propagation algorithmic rule is employed for the development of fuzzy based mostly neural network. The steps concerned within the algorithmic rule area unit as follows: 1st, initialize weights of the connections with random values. Second for every unit calculate internet input price, output price and error rate. Third, to handle uncertainty for every node, certainty live (c) for every node is calculated. Supported the knowledge live the choice is formed. The network made consists of three layers specifically Associate in Nursing input layer, a hidden layer Associate in Nursing an output layer. Sample trained neural network consisting of nine input nodes, three hidden nodes and one output node is shown in Figure two. once a clot or blood occupies over seventy fifth of expanse of the lumen of Associate in Nursingartery then the expected result is also a prediction of death or cardiopathy consistent with medical tips i.e. R is generated with regard to the given set of input file.

In Bayesian network structure discoveries, a contingent probability is that the chance of some conclusion, C, given some evidence/observation, E, wherever a dependence relationship exists between C and E. This conditional relationship permits Associate in Nursing investigator to achieve likelihood data concerning either C or E with the legendary outcome of the opposite. Currently take into account a posh downside with n binary variables, wherever the relationships among them don't seem to be clear for predicting one category output variable. If all variables were connected employing a single joint distribution, the equivalent of all nodes being 1st level oldsters, the quantity of potential combos of variables would be up to (2n-1). This ends up in the requirement for a really great amount of information [9, 10]. If dependence relationships between these variables may be determined leading to freelance variables being removed, fewer nodes would be adjacent to the node of interest. This parent node removal results in a big reduction within the range of variable combos, thereby reducing the quantity of required knowledge. Moreover, variables that area unit directly conditional, to not the node of interest however to the oldsters of the node of interest (as nodes four and five area unit with relevancy node1 [20], may be connected, that permits for a additional sturdy system once managing missing knowledge points. This property of requiring less data supported pre-existing understanding of the system’s variable dependencies could be a major advantage of Bayesian Networks [10]. Some more theoretical underpinnings of the Bayesian approach for classification are addressed in [11] and [12].

A Bayesian Network (BN) could be a comparatively new tool that identifies probabilistic correlations so as to form predictions or assessments of sophistication membership. Whereas the independence assumption could appear as a simplifying one and would so result in less correct classification, this has not been true in several applications. Parenthetically, many datasets area unit classified in [13] victimization the naïve Bayesian classifier, call tree induction, instance based mostly learning [22], and rule induction. These strategies area unit compared showing the naïve classifier because the overall best technique. To use a Bayesian Network as a classifier, first, one should assume that knowledge correlation is cherish applied mathematics dependence [23].

The purpose of the k Nearest Neighbours (kNN) [14] algorithmic rule is to use a info during which the info purposes area unit separated into many separate categories to predict the classification of a replacement sample point. Suppose a bank includes a info of people’s details and their credit rating. These details would most likely be the person’s money characteristics resembling what proportion they earn, whether or not they own or rent a house, and so on, and would be wont to calculate the person’s credit rating [21]. However, the method for calculative the credit rating from the person’s details is kind of valuable, therefore the bank would love to search out a way to scale back this value. They realise that by the terribly nature of a credit rating, those that have similar money details would be similar credit ratings. Therefore, they might wish to be ready to use this existing info to predict a replacement customer’s credit rating, while not having to perform all the calculations. kNN [15] is wide utilized in comparison of datasets and offers smart linear accuracy. A Support Vector Machine (SVM) [16] could be a discriminative classifier formally outlined by a separating hyper plane. In alternative words, given tagged coaching knowledge (supervised learning), the algorithmic rule outputs Associate in Nursing best hyper plane that categorizes new examples. In 2 dimensional area this hyper plane [17] could be a line dividing a plane in 2 elements wherever in every category lay in either facet. this is often a really complicated classifier, however is mostly utilized in clinical illness classification, because it produces terribly high accuracy in cases wherever two category classification is required [18], like in our case wherever we'd like to envision if the person is tormented by a illness or not. Consequent section describes our projected hybrid classifier, followed by the results and their comparison with alternative normal classifiers.

3. PROPOSED HYBRID DISEASE DETECTION CLASSIFIER

The block diagram of the proposed classifier can be shown in figure 1 as follows,

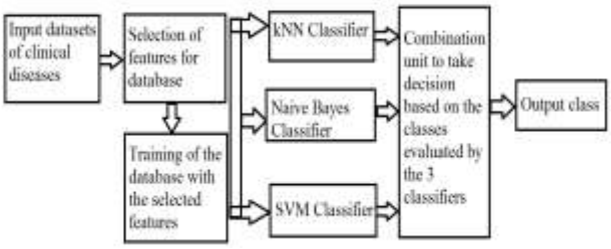

Input datasets of clinical diseases
Selection of features for database
Training of the database with the selected features
kNN Classifier
Naive Bayes Classifier
SVM Classifier
Combination unit to take decision based on the classes evaluated by the 3 classifiers
Output class

Figure1. Block diagram of the overall system

From the diagram, it can be seen that, the first block takes input from all the clinical disease datasets, and gives it to the feature selection block. This block basically collects all the datasets of various diseases, extracts the features from the datasets, divides the feature sets into training and testing sets, and then finally gives it to the feature selection block.

The feature selection block is an optional component of the system, in this block the variance of the features is calculated on per feature basis. These variances are then normalized in the range of 0 to 1, and then a mean value is calculated from these normalized variances. The features where the variance value is less than 20% of the mean value are discarded from the system, and the remaining feature sets are taken for further processing. This step helps in removing all the unwanted or redundant features from the dataset, and keeps only the valid or most varying features in the dataset. The output of this block is given to the database training block.

The database training block takes all the selected features and stores them in a local database along with the classes of the records. Each of these classes are further used for evaluation of the accuracy of the classifier under test.

In our proposed classification system, we are using 3 classifiers, namely kNN, Naive Bayes and SVM. kNN and Naive Bayes are already proven classifiers in previous work, in this work we are adding another classifier namely SVM and combining it with kNN and Naive Bayes in order to improve the overall accuracy of the system. The classifier works in the following manner,

- Classify the data from the kNN classifier and obtain a class C1
- Classify the data from Naive Bayes classifier and obtain the class C2
- Classify the data from SVM classifier and obtain the class C3
- If, C1 and C2 are equal, then the output class is C1
- If, C1 and C3 are equal, then the output class is C3
- If C2 and C3 are equal, then the output class is C2
- Else, the output class is C3

The classifier uses C3 as the default class if none of classes' match, due the fact that C3 is obtained from SVM and SVM is a very strong 2 class classifier, with good real time accuracy. In order to verify the results of SVM, we compare it's output class with the classes obtained from kNN and Naive Bayes, if all the classes are matching then the output is as per the class obtained from majority of the classifiers. This ensures that the system has good accuracy and takes minimal delay for result evaluation. We tested the results on 4 real time disease datasets taken from the UCI repository, and the results and it's analysis is discussed in the next section.

4. RESULTS AND ANALYSIS

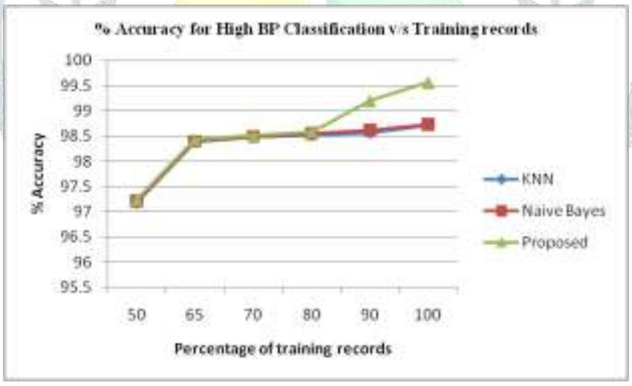
We compared our algorithm with kNN and Naive Bayes. kNN is a linear classifier, while Naive Bayes is a bayesian non-linear classifier. The results of both the classifiers when compared with our proposed approach is as shown in the following tables,

For High Blood Pressure, the Number of records were 2800, and the number of parameters per record were 30. We observed that the proposed classifier is 10% more accurate than the existing ones, the results of the same are as follows,

Table1. Delay and Accuracy of Proposed classifier for High BP dataset

Algo	KNN		Naive Bayes		Proposed	
Num Records	Delay (Sec)	Accuracy (%)	Delay (Sec)	Accuracy (%)	Delay (Sec)	Accuracy (%)
50	0.0024	97.1786	0.316	97.2143	0.0011	97.25
65	0.0025	98.3929	0.6838	98.3929	0.0011	98.4286
70	0.0028	98.48	0.7932	98.49	0.0012	98.52
80	0.0035	98.52	0.8854	98.55	0.0015	98.59
90	0.0038	98.57	0.9127	98.61	0.0018	99.21
100	0.0041	98.72	1.0225	98.73	0.0021	99.57

Figure2. Accuracy comparison for high BP dataset



For the Arrhythmia disease, there are 452 entries in the dataset, with each entry having 279 parameters. The results for the same can be shown as follow,

Table2. Comparison for Arrhythmia disease

Algo	KNN		Naive Bayes		Proposed	
Num Records	Delay (Sec)	Accuracy (%)	Delay (Sec)	Accuracy (%)	Delay (Sec)	Accuracy (%)
50	0.0028	85.3982	0.0076	85.177	0.0007	86.0619
65	0.0028	87.6106	0.0106	87.8319	0.0008	88.0531
70	0.0027	91.1504	0.0152	91.1504	0.0008	91.3717
80	0.0029	95.8752	0.0215	95.354	0.0008	95.7965
90	0.0031	96.9027	0.0251	96.6814	0.0008	97.1239
100	0.0032	98.8938	0.0311	98.8938	0.0008	99.115

For the Breast Cancer dataset, we had 116 records with 9 parameters per record. The results can be shown as follows,

Table 3. Comparison for breast cancer dataset

Algo	KNN		Naive Bayes		Proposed	
Num Records	Delay (Sec)	Accuracy (%)	Delay (Sec)	Accuracy (%)	Delay (Sec)	Accuracy (%)
50	0.0051	52.5862	0.0006	52.5862	0.0006	53.4483
65	0.0036	69.8276	0.0008	71.5517	0.0007	72.4138
70	0.0035	81.0345	0.0008	81.8966	0.0006	82.7586
80	0.0034	87.931	0.001	87.931	0.0006	88.7931
90	0.0038	93.9655	0.0011	93.9655	0.0006	94.8276
100	0.0035	100	0.0013	100	0.0006	100

While for the Parkinson's dataset, we have 195 records with 22 parameters in each record. The results can be observed as follows,

Table 4. Comparison for Parkinson's dataset

Algo	KNN		Naive Bayes		Proposed	
Num Records	Delay (Sec)	Accuracy (%)	Delay (Sec)	Accuracy (%)	Delay (Sec)	Accuracy (%)
50	0.0029	86.1538	0.0008	82.5641	0.0006	87.2
65	0.0028	89.7436	0.0011	86.1538	0.0006	90.5
70	0.0028	89.7236	0.0013	87.1795	0.0006	90.8
80	0.0028	89.7436	0.0016	87.1795	0.0006	91.5
90	0.0028	93.8462	0.0019	83.5897	0.0006	95.6
100	0.0028	100	0.0025	85.641	0.0006	100

The combined results for all the datasets can be as seen from the following table,

Table 5. Comparison of all algorithms

Algorithm	Mean Delay (ms)	Mean Accuracy (%)
kNN	0.0032	90.8437
Naïve Bayes	0.1975	89.3881
Hybrid	0.0009	91.5388

From the above tables we can observe that the hybrid algorithm is superior to the already existing algorithms, and can be used in real time with multiple datasets and produce the similar optimum results for each of the sets in real time.

5. CONCLUSION

From the obtained results, we can conclude that the proposed system is more than 10% accurate than the kNN and Naive Bayes based disease detection systems, and it is more than 15% faster in terms of delay of comparison when compared with similar techniques. Evaluation of the algorithms using other datasets also produces similar results, and thus can be used in real time systems.

6. FUTURE WORK

The proposed protocol demonstrates good accuracy for both stored and real time datasets. But it is observed that the system has moderate to high delay as the training set increases, thus it needs to be reduced by using machine learning optimizations, which can be carried out as a future work for this research.

7. ACKNOWLEDGEMENT

I am deeply indebted and would like to express my gratitude to my thesis guide Prof. Sapna Choudhary, Head P.G. Computer Science & Engineering, Shri Ram Group of Institution (SRGI), Jabalpur for his erudite guidance, affectionate encouragement and whole hearted involvement in my dissertation, without which it would have been difficult for me to complete this work. I express my thanks to all the faculty members and the staff of the Department of Computer Science & Engineering for providing me with the necessary facilities for completing my dissertation work. Special thanks to my husband for supporting me in my dissertation work and also throughout the curriculum. I express my heartfelt gratitude to my friends and also to each and every individual who was associated with my dissertation work, including those whom I may have inadvertently failed to mention. Finally, I am highly obliged to all my family members for their support and blessings.

RERERENCES

[1] E. Mirkes, KNN and Potential Energy (Applet). University of Leicester. Available: http: //www.math.le.ac.uk/people/ag153/homepage/KNN/KNN3.html, 2011.

[2] L. Kozma, k Nearest Neighbours Algorithm. Helsinki University of Technology, Available: http://www.lkozma.net/knn2.pdf, 2008

[3] N. Bhatia et al, Survey of Nearest Neighbor Techniques. International Journal of Computer Science and Information Security, Vol. 8, No. 2, 2010.

[4] F. Anguilli, Fast Condensed Nearest Neighbor Rule. Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005.

[5] . Frawley and Piatetsky-Shapiro, 1996. Knowledge Discovery in Databases:An Overview. The AAAI/MIT Press, Menlo Park, C.A.

[6]. Miller, A., B. Blott and T. Hames, 1992. Review of neural network applications in medical imaging and signal processing. Med. Biol. Engg. Comp., 30: 449-464.

[7]. Chen, J., Greiner, R.: Comparing Bayesian Network Classifiers. In Proc. of UAI-99, pp.101–108 ,1999.

[8]. Glymour, C., D. Madigan, D. Pregidon and P.Smyth, 1996. Statistical inference and data mining. Communication of the ACM, pp: 35-41.

[9]. Chen, J., Greiner, R.: Comparing Bayesian Network Classifiers. In Proc. of UAI-99, pp.101–108 ,1999.

[10]. "Hospitalization for Heart Attack, Stroke, or Congestive Heart Failure among Persons with Diabetes", Special report: 2001 – 2003, New Mexico.

[11]. “Heart disease” from http://wikipedia.org 8. Rumelhart, D.E., McClelland, J.L., and the PDF Research Group (1986), Parallel Distributed Processing, MA: MIT Press, Cambridge. 1994.

[12]. Heckerman, D., A Tutorial on Learning With Bayesian Networks. 1995, Microsoft Research.

[13]. Neapolitan, R., Learning Bayesian Networks. 2004, London: Pearson Printice Hall.

[14]. Krishnapuram, B., et al., A Bayesian approach to joint feature selection and classifier design.Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2004. 6(9): p. 1105-1111.

[15]. Shantakumar B.Patil, Y.S.Kumaraswamy, Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450- 216X Vol.31 No.4 (2009), pp.642-656 © EuroJournals Publishing, Inc. 2009.

[16]. Sellappan Palaniappan, Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, 978-1-4244-1968-5/08/$25.00 ©2008 IEEE.

[17]. Pedro Domingos , Michael Pazzani , On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, Machine Learning, 29, 103–130 (1997) c° 1997 Kluwer Academic Publishers. Manufactured in The Netherlands.

[18]. Richard N. Fogoros, M.D, The 9 Factors that Predict Heart Attack 90% of heart attacks are determined by these modifiable risk factors, About.com Guide.

[19]. Harleen Kaur and Siri Krishan Wasan, Empirical Study on Applications of Data Mining Techniques in Healthcare, Journal of Computer Science 2 (2): 194-200, 2006 ISSN 1549-3636 © 2006 Science Publications.

[20]. Bressan, M. and J. Vitria, On the selection and classification of independent features. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2003. 25(10): p. 1312-1317.

[21]. Domingos, P. and M. Pazzani, On the optimality of the simple Bayesian classifier under zeroone loss. Machine Learning, 1997. 29(2-3): p. 103-30.

[22]. Juan Bernabé Moreno, One Dependence Augmented Naive Bayes, University of Granada, Department of Computer Science and Artificial Intelligence.

[23]. Giorgio Corani, Marco Zaffalon, JNCC2: The Java Implementation Of Naive Credal Classifier 2, Journal of Machine Learning Research 9 (2008) 2695-2698