Feature Selection Using FCBF in Type II Diabetes Databases

Sarojini Balakrishnan
Department of Computer Applications
K.L.N. College of Information Technology
Madurai, India
balakrishnan.sarojini@gmail.com

and

Ramaraj Narayanaswamy
Department of Computer Science & Engineering
G.K.M. College of Engineering & Technology
Chennai, India
ramaraj_tce@yahoo.co.in

***Abstract*— Data mining techniques have been widely applied to extract knowledge from medical databases. Data mining is the search for relationships and global patterns that exist in large databases but are `hidden' among the vast amount of data, such as a relationship between patient data and their medical diagnosis. Usually medical databases are high dimensional in nature. If a training dataset contains irrelevant and redundant features (i.e., attributes), classification analysis may produce less accurate results. In order for data mining algorithms to perform efficiently and effectively on high-dimensional data, it is imperative to remove irrelevant and redundant features. Data pre-processing is required to prepare the data for data mining tasks to increase the predictive accuracy. Feature selection is one of the important and frequently used data preprocessing technique for data mining applications in medicine. This paper illustrates, the application of feature selection technique in medical databases, will enable to find small number of informative features leading to potential improvement in medical diagnosis. It is proposed to find an optimal feature subset of the PIMA Indian Diabetes Dataset using Symmetrical Uncertainty Attribute set Evaluator and Fast Correlation-Based Filter (FCBF). The approach is validated by means of performance enhancement of the Libsvm classifier.**

***Keywords*— Data mining, Feature selection,data preprocessing, predictive accuracy,Symmeterical Uncertainty Attribute set evaluator, FCBF, Libsvm.**

I. INTRODUCTION

Medical Data mining is the search for relationships and patterns within the medical data that could provide useful knowledge for effective medical diagnosis. Extracting knowledge from these health care databases can lead to discovery of trends and rules for later diagnostic tools. Consequently, the predictability of disease will become more effective and early detection of disease will aid in increased exposure to required patient care and improved cure rates [1]. Data mining methods have been applied to a variety of medical domains to improve medical diagnosis [2]. Some include predicting breast cancer survivability using data mining techniques [3], application of data mining to discover subtle factors affecting the success, failure of back surgery which led to improvements in care [4], data

mining classification techniques for medical diagnosis decision support in a clinical setting [5] and the techniques of data mining used to search for relationships in a large clinical database [6].

Data from medical sources are voluminous and are high dimensional. The medical data are characterized by their incompleteness (missing parameter values), incorrectness (noise in the data), sparseness (few and/or non-representable patient records are available) and inexactness (inappropriate selection of parameters for a given task). In a real-world environment, there are many possible reasons why the inaccurate or inconsistent data occur in a medical database, e.g., equipment mal functioning, the deletion of data instances (or records) due to the inconsistency with other recorded data, not entering data due to misunderstanding, considering the data as an unimportant at the time of entry, etc. These data characteristics need to be considered in the design of analysis tools for prediction, intelligent alarming and therapy support [7]. Much research work has focused on the development of data mining algorithms that can learn regularities in these rich, mixed media data. Many factors affect the success of data mining on medical datasets. If information is irrelevant or redundant or the data is noisy and unreliable then knowledge discovery during training is more difficult.

Feature selection is the process of identifying and removing as much of the irrelevant and redundant information as possible [8]. Feature selection is often considered as a necessary preprocess step to analyze these data, as this method can reduce the dimensionality of the datasets and often conducts to better analysis [9]. Research [10] shows that the reasons for feature selection include improvement in performance prediction, reduction in computational requirements, reduction in data storage requirements, reduction in the cost of future measurements and improvement in data or model understanding. Much research work in data mining has gone into improving the predictive accuracy of the classifiers by applying the techniques of feature selection. Feature selection techniques identify the features that mostly improve the predictive accuracy of the classifiers. Many authors have reported improvement in the performance of the classifier when feature selection algorithms are used [11,12,13].

Two models of feature selection exist depending on whether the selection is coupled with a learning scheme or not [14,15]. The filter model carries out the feature subset selection independent of the learning algorithm and the wrapper model carries out the feature subset selection using a learning algorithm to measure the classification accuracy. In this research, we propose a filter based feature selection approach Symmetrical Uncertainty Attribute set selector and FCBF (Fast Correlation-Based Filter) search to remove both irrelevant and redundant features. FCBF uses correlation measures for relevant and redundant analysis [16]. From the remaining features, we empirically evaluate the classification effectiveness of Libsvm Classifier [17] on the reduced feature set of Pima Indian Diabetes [18] data set.

II. RELATED WORK

Medical applications of data mining include prediction of the effectiveness of surgical procedures, medical tests and medications, and discovery of relationships among clinical and pathological data [19]. Moustakis and Charissis' work [20] surveyed the role of machine learning in medical decision making and provided an extensive literature review on various ML applications in medicine that could be useful to practitioners interested in applying ML methods to improve the efficiency and quality of decision making systems in medical applications. Data mining techniques have been applied to a variety of medical domains to improve medical decision making [2]. A comparison of

different learning models used in Medical Data Mining and a practical guideline how to select the most suited algorithm for a specific medical application is found in [7].

The importance of feature selection in medical domain is found in [21]. Feature selection has been an active and fruitful field of research and development for decades in statistical pattern recognition [22], machine learning [23,24], data mining [25] and statistics [26]. It has proven in both theory and practice effective in enhancing learning efficiency, increasing predictive accuracy, and reducing complexity of learned results [27,28,29]. In the presence of hundreds or thousands of features, researchers notice [30,31] that it is common that a large number of features are not informative because they are either irrelevant or redundant with respect to the class concept. Feature selection has found success in many applications like text categorization [30], image retrieval [32], genomic microarray analysis [31], customer relationship management [33], and intrusion detection [34].

Filter methods can be further categorized into two groups, namely attribute evaluation algorithms and subset evaluation algorithms, based on whether they rate the relevance of individual features or feature subsets [29]. Attribute evaluation algorithms rank the features individually and assign a weight to each feature according to each feature's degree of relevance to the target feature. Yu and Liu (2003) [16] note that attribute evaluation methods are likely to yield subsets with redundant features since these methods do not measure the correlation between features. Subset evaluation methods, in contrast, select feature subsets and rank them based on certain evaluation criteria and hence are more efficient in removing redundant features. A number of experimental studies [29,14,35] have shown that irrelevant and redundant features can dramatically reduce the predictive accuracy of models built from data. In the filter approach, examples of evaluation functions are probabilistic distance, interclass distance, information-theoretic or probabilistic dependence measures [36]. These measures are often considered as intrinsic properties of the data, because they are calculated directly on the raw data instead of requiring learning model that smoothes distributions or reduces the noise [37].

Feature selection is one effective means to remove irrelevant features [29]. Correlation is widely used in machine learning for relevance analysis [16]. Researchers often resort to various approximations to determine relevant features (e.g., relevance is determined by correlation between individual features and the class) [38,16]. However, a single feature can be considered irrelevant based on its correlation with the class; but when combined with other features, it becomes very relevant. [28,14, 38] in their works reported the existence and effect of feature redundancy.

Supervised learning systems such as classification have been successfully applied in a number of medical domains, for example, in localization of a primary tumor, prognostics of recurrence of breast cancer, diagnosis of thyroid diseases, and rheumatology [39].

The Support Vector Machine (SVM) was originally designed for binary classification problems [40]. A support vector machine (SVM) [41] is a popular and much applied supervised machine learning method. Support Vector Machines (SVMs),[42,43] one of the most actively developed classifiers in the machine learning community, have been successfully applied to a number of medical problems [44, 45,46,47,48].

For supervised learning, the primary goal of classification is to maximize predictive accuracy; therefore, predictive accuracy is generally accepted and widely used as the primary measure by researchers and practitioners [35]. The performance of a classifier can also be visualized by using a Receiver Operating Characteristic (ROC)

curve [4]. The performance of the classifier before and after feature selection can be evaluated based on the Area under the ROC curve, abbreviated **AUC** [1].

III. DATASET

The experiments were performed on the Pima Indian diabetes dataset from the UCI (University of California at Irvine) machine-learning repository [18], which consists of 768 complete instances described by 8 features and a predictor class. The class value 1 interpreted as "tested positive for diabetes" is found in 268 numbers of instances and class value 0 in 500 numbers of instances. There is no missing data present in the training dataset.

IV. OBJECTIVE/PROBLEM DEFINITION

The objective of this research is to derive the optimal feature subset for the PIMA dataset that improves the performance of the Libsvm classifier. It is advantageous to limit the number of input features in a classifier in order to have a good predictive and less computationally intensive model. With a small feature set, the explanation of rationale for the classification decision can be more easily realized. In the area of medical diagnosis, a small feature subset means fewer test and less diagnosis costs.

We propose to eliminate redundant and irrelevant features of the PIMA diabetes dataset to improve the predictive performance of the Libsvm classifier. We use Symmetrical Uncertainty (SU) with FCBF search method to evaluate the worth of a set of features by measuring the symmetrical uncertainty with respect to another set of features and to remove redundant and relevant features. The algorithm work as follows.

Given a N samples of data set with N features and a class C, the feature selection problem is to find from the M-dimensional observation space, S^M, a subspace of m features, S^m. The total number of subspaces is $2^{M.}$

The feature selection process consists of two phases: In the first phase Symmetrical Uncertainity, SUi,c, the measure of correlation between the feature i and the class C, is calculated for each feature. The features whose SU value greater than a threshold value δ are chosen to form a feature subset S_{best}, with the vales ordered in the descending order of SU values.

In the second phase, S_{best} is further processed remove redundant features. Start with the first feature to eliminate all features that are redundant to it and keep only the predominant ones among all the selected relevant features.

Fi is chosen if fi is SU(Fi,C)≥SU(Fj,C)and SU(Fi,Fj)≥ SU(Fj, C)

The symmetrical uncertainty is calculated using the formula [50],

$$SU(X,Y) = 2\left[\frac{IG(X|Y)}{H(X)+H(Y)}\right]$$

Where IG(X/Y) is the measure of Information Gain [51],

$$IG(X|Y) = H(X) - H(X|Y).$$

And

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)),$$

H(X) is the Entropy of a variable X

$$H(X|Y) = -\sum_j P(y_j)\sum_i P(x_i|y_j)\log_2(P(x_i|y_j)),$$

H (X/Y) is the Entropy of X after observing values of Y.

V. EMPIRICAL STUDY

The objective of this section is to evaluate the proposed approach by means of various performance indicators.

Experimental set up:

We used the Libsvm classifier and symmetrical uncertainty Attribute set Evaluator and FCBF implemented in the machine learning library with Java implementation "WEKA 3.5.2" [52] for our experiments. The experiments are performed on the PIMA dataset.

To implement our proposed approach, we used the RBF kernel function for the Libsvm classifier as the RBF kernel function can analyze higher-dimensional data and requires that only two parameters, C and γ to be defined [53]. Research [54] shows that setting proper model parameters can improve the classification accuracy of SVM. The grid search approach [53] is used to find the best C and γ.

The classification accuracy is evaluated using 10-fold cross validation test. Cross-validation involves breaking a dataset into 10 pieces, and on each piece, testing the performance of a predictor build from the remaining 90% of the data. The classification accuracy was taken as the average of the 10 predictive accuracy values.

The performance of the proposed approach is analyzed using two criteria: the accuracy of the classifier and the area under the ROC. The accuracy of the classifier is calculated using the formula [55].

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Where TP: the number of true positives (number of 'YES' patients predicted correctly), TN: the number of true negatives (number of 'NO' patients predicted correctly), FP: the number of false positives (number of 'YES' patients predicted as 'NO') and FN: the number of false negatives (number of 'NO' patients predicted as 'YES')

The performance of a classifier is visualized by using a Receiver Operating Characteristic (ROC) curve. The area under the ROC curve (AUC) is used as a tool for comparing the performance of the classifier on the removal of each feature. The curve that has a larger AUC is better than the one that has a smaller AUC.

Experimental Results and Discussion:

The values of best C and γ obtained for the PIMA dataset using grid search is 12.0 and 0.1 respectively.

The irrelevant and non-redundant features are removed using Symmetrical Uncertainty Attributeset Evaluator and FCBF search and the selected features are ranked in the decreasing order of SU values. The experimental results show that the features Glucose tolerance test, body mass function, age and diabetes pedigree function are selected as highly relevant and non-redundant features. These features are used for the classification purpose.

Table – I shows the performance of the classifier before and after feature selection. The performance of the classifier is evaluated based on the Accuracy of the classifier and the Area Under ROC Curve.

TABLE 1
PERFORMANCE OF THE CLASSIFIER BEFORE AND AFTER FEATURE SELECTION

Features Used	True Positive	False Positive	True Negative	False Negative	No.of correctly classified instances	No.of Incorrectly classified instances	Accuracy
All features	152	116	443	57	595	173	77.474
4	156	112	443	57	599	169	77.9948

The experimental results show that the accuracy of the classifier has improved with the removal of the irrelevant and redundant features.

Another way of evaluating the performance of a classifier is by the analysis of the ROC curve. The area under ROC for the whole set of features is 0.8344 and for the optimal feature subset it is 0.83. The results show that we could get the same AUC for the whole set of feature sets and for the reduced subset of feature. It is depicted in Figure 1 and Figure-2.

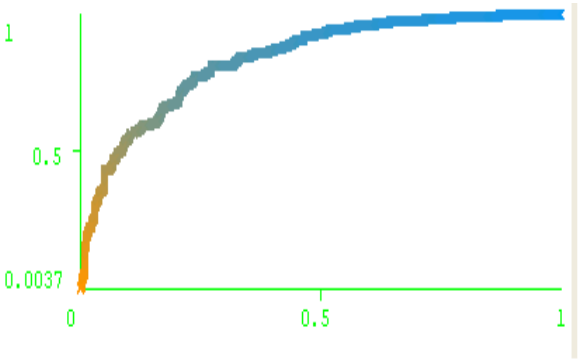

Figure 1:ROC before feature selection

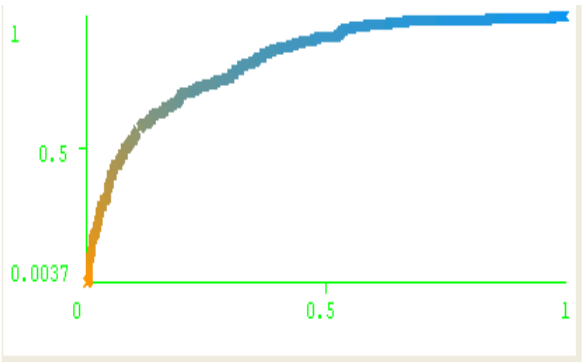

Figure 2: ROC after feature selection

VI. CONCLUSION

Many predictive data mining methods have been successfully applied to a variety of practical problems in medical domain. This research work attempts to emphasize feature selection techniques in clinical decision support tools could empower the medical community to improve the quality of diagnosis through the use of technology.

The proposed Symmetrical Uncertainty and FCBF feature selection approach produces a feature reduction of 62.5%. The accuracy of the classifier slightly improves. In medical domain, it is desirable the feature selection improves classification accuracy, but it is appreciable with reduced number of features if we could get the same classification accuracy as the whole set of features as reduction in the features means diagnosis with less number of tests the patient should undergo.

The approach is simple and effective and augments the argument simple methodologies are better for medical data mining.

REFERENCES

[1] Roshawnna Scales, Mark Embrechts, “Computational intelligence techniques for medical diagnostics”.

[2] Kononeko, I.,Kukar, M.(1995) Machine learning for medical diagnosis. *Workshop on Computer-Aided Data Analysis in Medicine*, CADAM-95, IJS Scientific Publishing, Ljubljana.

[3] Dursun Delen*, Glenn Walker, Amit Kadam ”Predicting breast cancer survivability: a comparison of three data mining methods” Artificial Intelligence in Medicine doi:10.1016/j.artmed.2004.07.002.

[4] Hedberg, SR. (1995) The data gold rush. Byte, 1995;Oct :83-88.

[5] Herron, “Machine Learning for Medical Decision Support: Evaluating Diagnostic Performance of Machine Learning classification Algorithms”

[6] Prather J. C., Lobach D. F., Goodwin L. K., Hales J. W.,Hage M. L., Edward Hammond W., “Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse”, 1997.

[7] Lavrac N. “Selected techniques for data mining in medicine” Artif Intell Med 1999; 16:3—23.

[8] Liu H., Sentino R, "Some issues on scalable Feature Selection, Expert Systems with Application", vol 15, pp 333-339, 1998.

[9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," Machine Learning, vol. 46, no. 1-3, pp. 389–422, 2002. [Online]. Available: citeseer.ist.psu.edu/guyon02gene.html

[10] T.Guyon, "Practical Correlation: From Correlation to casuality",

[11] Almuallim, H., and Dietterich, T.G., Efficient algorithms for identifying relevant features, In Proceedings of the Ninth Canadian Conference on Artificial Intelligence, Vancouver, BC: Morgan Kaufmann, 1992.

[12] Aha, D.W., and Bankert, R. L., A comparative evaluation of sequential feature selection algorithms, In D. Fisher & J.-H. Lenz (Eds.), Artificial Intelligence and Statistics V. New York: Springer-Verlag. 1996.

[13] W Siedlecki and J. Skalansky, On automatic feature selection, Int. J. Pattern Recog. Art. Intell. vol. 2, no.2.p~1.9 7-220. 1988.

[14] Kohavi, R., & John, G. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97, 273–324.

[15] Das, S. (2001). Filters, wrappers and a boosting-based hybrid for feature selection. Proceedings of the Eighteenth International Conference on Machine Learning (pp. 74–81).

[16] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proc 12th Int Conf on Machine Learning (ICML-03), Washington, D.C.*, pages 856–863, San Francisco, CA, 2003. Morgan Kaufmann.

[17] Yasser EL-Manzalawy and Vasant Honavar, WLSVM : Integrating LibSVM into Weka Environment, 2005. Software available at http://www.cs.iastate.edu/~yasser/wlsvm

[18] C.L. Blake, C.J. Merz, UCI repository of machine learning databases, Website: http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998

[19] Ranjit Abraham, Jay B.Simha, Iyengar S. "Medical datamining with a new algorithm for feature selection and Naïve Bayesian classifier", 10th International Conference on Information Technology.

[20] Moustakis, V. and Charissis, G. (1999). Machine learning and medical decision making, In Proceedings of Workshop on Machine Learning in Medical Applications, Advance Course in Artificial Intelligence- ACAI99, Chania, Greece, 1-19.

[21] I. Kononenko, I. Bratko, and M. Kukar. Application of machine learning to medical diagnosis. In Machine Learning and Data Mining: Methods and Applications, pages 389-408. John Wiley & Sons

[22] P. Mitra, C. A. Murthy, and S. K. Pal. Unsupervised feature selection using feature similarity, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3):301–312, 2002.

[23] H. Liu, H. Motoda, and L. Yu,. Feature selection with selective sampling, In Proceedings of the Nineteenth International Conference on Machine Learning, pages 395–402, 2002b.

[24] M. Robnik-Sikonja and I. Kononenko., Theoretical and empirical analysis of Relief and ReliefF. Machine Learning, 53:23–69, 2003.

[25] M. Dash, K. Choi, P. Scheuermann, and H. Liu., Feature selection for clustering – a filter solution, In Proceedings of the Second International Conference on Data Mining, pages 115–122, 2002.

[26] A. Miller., Subset Selection in Regression, Chapman & Hall/CRC, 2 edition, 2002.

[27] H. Almuallim and T. G. Dietterich., Learning boolean concepts in the presence of many irrelevant features, Artificial Intelligence, 69(1-2):279–305, 1994.

[28] D. Koller and M. Sahami., Toward optimal feature selection, In Proceedings of the Thirteenth International Conference on Machine Learning, pages 284–292, 1996.

[29] A. L. Blum and P. Langley., Selection of relevant features and examples in machine learning, Artificial Intelligence, 97:245–271, 1997.

[30] Yang, Y., & Pederson, J. O. (1997). A comparative study on feature selection in text categorization. Proceedings of the Fourteenth International Conference on Machine Learning (pp. 412–420).

[31] Xing, E., Jordan, M., & Karp, R. (2001). Feature selection for high-dimensional genomic microarray data. Proceedings of the Eighteenth International Conference on Machine Learning (pp. 601–608).

[32] D.L.Swets and J.J.Weng,"Efficient content-based image retrieval using automatic feature selection", In IEEE International Symposium on Computer Vision, pages 85-90, 1995

[33] K.S.Ng and H.Liu," Customer retention via data mining",AI review, 14(6):569-590, 2000

[34] W.Lee, S.J.Stofolo, and K.W.Mok, "Adaptive Intrusion detection: A data mining approach.AI Review,14(6):533-567,2000.

[35] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant feature and the subset selection problem. In *Proceedings of the Eleventh International Conference on Machine Learning*, pages 121–129, 1994.

[36] M. Dash and H. Liu. Feature selection for classification. Intelligent Data Analysis, IOS Press 1997.

[37] PATRICK E. MEYER, COLAS SCHRETTER AND GIANLUCA BONTEMPI,"Information-theoretic feature selection in microarray data using variable complementarity", http://www.ulb.ac.be/di/mlg/

[38] M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 359–366, 2000.

[39] Richards G, Rayward-Smith VJ, Sonksen PH, Carey S, Weng C. "Data mining for indicators of early mortality in a database of clinical records", Artif Intell Med 2001; 22:215—31.

[40] R. Bellazzi, B. Zupan, Predictive data mining in clinical medicine: Current issues and guidelines, *Int.J.Med.Inform.(2006),doi:10.1016/j.ijmedinf.2006.11.006*

[41] B.Boser, I.Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *5th Annual ACM Workshop on COLT*, pages 144-152, 1992.

[42] V. Vapnik, *The Nature of Statistical Learning Theory*. NewYork: Springer Verlag, 1995

[43] C.Cortes and V.Vapnik, "Support vector networks," *Mach. Learning*, vol. 20, pp 273–297, 1995.

[44] K. Takeuchi and N. Collier, "Bio-medical entity extraction using support vector machines," *Artif. Intell. Med.*, vol. 33, pp. 125–137, 2005.

[45] G. Cohen, M. Hilario,H. Sax, S.Hugonnet, andA.Geissbuhler, "Learning from imbalanced data in surveillance of nosocomial infection," *Artif. Intell. Med.*, vol. 37, pp. 7–18, 2006.

[46] M. E. Mavroforakis, H. V. Georgiou, N. Dimitropuoulos, D. Cavouras, and S. Theodoridis, "Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector machine classifiers," *Artif. Intell. Med.*, vol. 37, pp. 145–162, 2006.

[47] Arodz,M.Kurdziel, E. O. D. Sevre, and D.A.Yuen, "Pattern recognition techniques for automatic detection of suspicious-looking anomalies in mammograms," *Comput. Methods Programs Biomed.*, vol. 79, pp. 135–149, 2005.

[48] L. Ramirez, N. G. Durdle, V. J. Raso, and D. L. Hill, "A support vector machines classifier to assess the severity of idiopathic scoliosis from surface topology," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 1, pp. 84–91, Jan. 2006.

[49] A. Swets, R. M. Dawes, and J. Monahan. "Better decisions through science", Scientific American, 283:82–87, October 2000.

[50] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, 1988.

[51] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[52] I. H. Witten and E. Frank. *Data Mining - Pracitcal Machine Learning Tools and Techniques with JAVA Implementations*. Morgan Kaufmann Publishers, 2000.

[53] Chen, Y.-W., & Lin, C.-J. (2005). Combining SVMs with various feature selection strategies. Available from http://www.csie.ntu.edu.tw/~cjlin/papers/features.pdf.

[54] Cheng-Lung Huang, Hung-Chang Liao b, Mu-Chen Chen c, "Prediction model building and feature selection with support vector machines in breast cancer diagnosis ", Expert Systems with Applications", 2008, 578-587 doi:10.1016/j.eswa.2006.09.041

[55] Rayner Alfred, "Knowledge Discovery: Enhancing Data Mining and Decision Support Integration"