View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Federation ResearchOnline

AWSum - Applying Data Mining in a Health Care Scenario

Anthony Quinn [2] , Herbert F. Jelinek [1], Andrew Stranieri [2], John Yearwood [2]
[1] *School of Community Health, Charles Sturt University*
Albury Wodonga, hjelinek@csu.edu.au
[2] *Information Technology and Mathematical Sciences, University of Ballarat*
Gear Ave , Mt Helen 3350, quinn@clearmail.com.au

Abstract

*This paper investigates the application of a new data mining algorithm called **A**utomated **W**eighted **Sum**, (AWSum), to diabetes screening data to explore its use in providing researchers with new insight into the disease and secondarily to explore the potential the algorithm has for the generation of prognostic models for clinical use.*

There are many data mining classifiers that produce high levels of predictive accuracy but their application to health research and clinical applications is limited because they are complex, produce results that are difficult to interpret and are difficult to integrate with current knowledge and practises. This is because most focus on accuracy at the expense of informing the user as to the influences that lead to their classification results. By providing this information on influences a researcher can be pointed to new potentially interesting avenues for investigation. AWSum measures influence by calculating a weight for each feature value that represents its influence on a class value relative to other class values.

The results produced, although on limited data, indicated the approach has potential uses for research and has some characteristics that may be useful in the future development of prognostic models.

1. Introduction

In practice, many data mining exercises using data drawn from patients with particular conditions are performed to provide medical researchers with some insight into the disease that could lead to a greater understanding of the condition and suggest possible interesting directions for research. In addition, the use of data mining findings has the potential to inform the development of diagnostic or prognostic models for use in clinical practice, though this is challenging.

Wyatt and Altman [14] highlight the lack of uptake of diagnostic models by clinicians with a few exceptions such as the Glasgow Coma Scale [8]. The reasons given for this include a lack of adequate evidence of credibility, accuracy, generality and effectiveness. By implication the medical researcher must also be wary of these concerns when presenting findings that have application in the clinical field.

Most of the statistical and data mining techniques currently available are complex and have many of the shortcomings outlined by Wyatt and Altman. This limits their usefulness as research tools and also in suggesting directions for the development of clinical models.

Decision trees and rules such as those generated by C4.5 [4] can be difficult to interpret and rules in some cases need to be applied in a specific order. The pruning of trees or rules to make them manageable may remove relevant factors simply because other factors are more relevant for the classification. Complexity is also introduced in some tree algorithms by the discretisation of continuous features based on the separation of the class value rather than medical knowledge.

Bayesian approaches [2] can also be problematic in that probabilities at many nodes affect the classification requiring some amount of reverse engineering to determine the influences on the classification. The selection of important features is also an issue because often a scoring metric based on an improvement in classification rather than relevance is used.

Other techniques such as Neural Networks [6] and Support Vector Machines [9] fall into the category of ”black box”. By this we mean that the influences that lead to the classification are not obvious or transparent and this limits the researchers ability to gain an understanding of the problem domain from the classifier.

Of the many statistical approaches available, forms of logistic regression are currently popularly used in many medical applications. Although these have solid theoretical underpinnings Wyatt and Altman [14] found that in as many as one in five statistical models the underlying assumptions were violated affecting the integrity of the approach. Another difficulty with statistical models when used for research is that their usefulness is limited when only a small amount of data is available. This could be seen as a positive in a clinical model because we would want to be sure of the basis of the model but can be a shortcoming in research. In medical research often only small data samples are available and it is the rare item that is being searched for.

The classifier investigated in this paper, **A**utomated **W**eighted **Sum** (AWSum) [5]. maintains a comparable classification accuracy while presenting the user with information that is simple to interpret. It does this by formulating a weight for each feature value that indicates its relative influence on the outcome. Thus the user can determine the important factors influencing the outcome as well as the weight that should be

given to these factors. An important part of AWSum's approach is the inclusion of the expert's knowledge in assessing the validity of the information and so any anomalies in the influence weights are investigated as points of interest that may potentially expand the expert's knowledge or at the least require further explanation. The application of AWSum to a diabetes database supplied by Charles Sturt University [12] is investigated demonstrating potentially valuable insights that can be obtained using this approach. The accuracy and value of these insights has been assessed by one of the authors..

AWSum establishes its influence and weights by considering the strength of the associations between the feature values and class values. This differs from other approaches in that it is an explicit focus on feature values rather than features, which is the case in most classifiers. The intuition behind AWSum's approach is that each feature value has an influence on the classification that can be represented as a weight and that combining these influence weights gives an influence score for an example. This score can then be compared to a threshold in order to classify the example. This can be seen as a weighted voting system or a combination of evidence approach that is similar to the methodology applied in the clinician - patient scenario. Thus the approach outlined here is simple and meets the criteria of transparency and ease of application.

The algorithm for calculating and combining weights, and determining thresholds is briefly described in section 2.

2. THE ALGORITHM

The following section briefly describes the algorithm. It consists of 2 steps; the first is to calculate the influence weights and the second to classify new examples

A. Influence weights

The first phase of the AWSum approach lays the foundations for classification by calculating influence weights for each feature value. Calculating the conditional probability of the outcome given the feature value provides the level of association between the feature value and the outcome. To calculate an influence weight the level of association for each class value, for a given feature value, needs to be combined into a single figure. We will first consider the case of a binary classifier.

A feature value's influence weight, *W* represents its influence on each class value and so it needs to simultaneously represent the feature value's association with both class values. To achieve this one association is considered positive and the other negative. This leads to a range for the influence weight of -1 to 1, where a certainty of one class value produces a weight of -1 and a certainty of the other class value a weight of 1. By summing the two associations we arrive at a single influence weight that represents the feature value's influence on both class values simultaneously. Equation 1 demonstrates this calculation and figure 1 shows an example where $Pr\left(O_1|Fv\right) = 0.2$, or -0.2 when mapped and $Pr\left(O_2|Fv\right) = 0.8$.

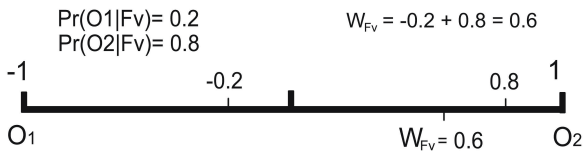

Fig. 1: Binary class example

$$W = Pr\left(O_1|Fv\right) + Pr\left(O_2|Fv\right) \tag{1}$$

Additional assumptions are required to be made in the case of class features that are ternary or of a higher order. This is discussed below in Section 4.

B. Classification

Classification of an example is achieved by combining the influences of the weights for each of the example's feature values into a single score. By summing and averaging influence weights we are able to arrive at a scaled score that represents a combination of the evidence that the example belongs to one class and not to another. Equation 2 depicts this. Performing the combination by summing and averaging assumes each feature value's influence is equally comparable. Although this is a relatively naive approach, it is quite robust as described later in this section. It also leaves open the possibility of using other functions for the combining of influence weights, much the same as different kernel functions can be used in support vector machines.

$$e_1 = \frac{1}{n}\sum_{m=1}^{n} W_m \tag{2}$$

e_1 = the influence weight of the i^{th} example
n = the number of features

The influence score for an example is compared to threshold values that divide the influence range into as many segments as there are class values. For instance, a single threshold value is required for a binary classification problem so that examples with an influence score above the threshold are classified as one class value, and those with a score below the threshold are classified as the other class value. Each threshold value is calculated from the training set by ordering the examples by their influence weight and deploying a search algorithm based on minimising the number of incorrect classifications. For instance, the examples with total influence scores that fall to the left of the threshold in Figure 2 are classified as class outcome *A*. This however includes two examples that belong to class *B* in the training set and so these two examples are also misclassified but the number of misclassifications have been minimised. Two examples to the right of the threshold are misclassified as class *B* when they are *A's*. In cases where there are equal numbers of correctly and incorrectly classified examples the threshold is placed at the mid-point under the assumption that misclassification of class *A* and *B* is of equal cost.

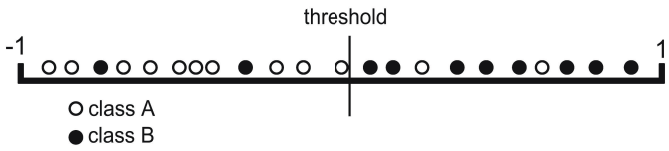

Fig. 2: Threshold optimisation

New examples can be classified by comparing the example's influence score to the thresholds. The example belongs to the class in which its influence score falls.

AWSum is suited to nominal feature values and class outcomes although it is not necessary that they are ordinal. Continuous numeric features require discretisation before use in AWSum. While there is a potential for developing a distinct method of discretisation in AWSum the research to date has used Fayyad and Irani's MDL method [13].

This method of discretisation could be seen as somewhat arbitrary from a medical perspective rather than based on accepted medical knowledge but because we were directing the algorithm at research we were interested to see the level of concurrence between accepted threshold and automated selection techniques.

C. Combinations of Feature Values

The combining of influence weights for single feature values into a total influence score for an example and using this to classify is intuitively based. However, it is plausible that feature values may not individually be strong influences on a class outcome but when they occur together the combination is a strong influence. For example both *drug A* and *drug B* may individually be influential toward low blood pressure but taken together lead to an adverse reaction that results in exceedingly high blood pressure.

The influence weights for each feature value combination can be calculated in the same way as they were for the single feature values. These combinations of feature values can contribute to an increase in accuracy and provide insight. Analysts can use them to identify feature values that have interesting interactions. This is achieved by comparing the influence weights of the individual component feature values of the combination to the influence weight of the combination. If they are markedly different this indicates a level of interaction between the feature values.

D. Model selection

AWSum calculates an influence weight for each feature value and all combinations of feature values and so a comparison of the influence of the feature value combination to its parents is possible. By this we mean that a feature value combination containing two feature values can be compared with the feature value weight of each of the components that make it up. In doing so the difference between the influence weight of the parent and child can be calculated 3. If the influence can be attributed to a parent, or if the weight of the combination is not significantly different to the influence calculated for combining the two single feature influence weights using AWSum's averaging method 2 then there is no need to include the child in the classification model. This also leads to an ability to identify combinations of feature values that interact strongly in a way different to their constituent feature values which can provide insight into the data as discussed above.

$$W_{diff} = W_{F_1} - W_{F_1|F_2} \tag{3}$$

To select a model the combinations of feature values are ordered according to the magnitude of the influence weight difference. The first N combinations, where N ranges from 1 to the number of possible combinations, are added and N incremented until the classification is maximised.

This approach suffers from an intrinsic shortcoming inherent in stratified cross validation, being that the model may be different for each run of the classifier. This issue needs to be addressed by accessing more data and testing the models for convergence.

3. Experiments

Four datasets were sourced from the UCI Repository [[1]] for the comparative evaluation of the AWSum approach. In addition, the Diabetes (DM) dataset [12], with 77 features, 2 classes, 1930 instances, and many missing values, was used. Ten fold stratified cross validation was used in all experiments. Table 1 shows the classification accuracy by other techniques using the Weka [10] suite alongside results from AWSum. AWSum Single refers to the results using single feature values independently, without considering any interaction between feature values. AWSum Triples shows the classification accuracies achieved by including the influence weights for combinations of feature values up to a combination of three feature values. The Weka implementation of the following commonly used classifiers were used for comparison: Naive Bayesian Classifier NBC which uses conditional probability and an assumption of independence, Tree Augmented Bayesian Network TAN which includes important dependencies, C4.5 a well accepted tree based classifier that uses information gain to select nodes, Support Vector Machine SVM which is a geometric approach and Logistic Regression which is a well accepted statistical approach. Table1 illustrates that AWSum performs comparably on all datasets.

Table 1: Classifier comparison using single feature value influence weights only

Data	AWSum Single	AWSum Triple	NBC	TAN	C4.5	SVM	Log
Heart	83.14	**89.90**	84.48	81.51	78.87	84.16	84.48
Iris	94.00	94.00	94.00	94.00	96.00	**96.67**	93.33
Mush	95.77	99.37	95.83	99.82	**100**	**100**	**100**
Vote	86.00	**97.48**	90.11	94.25	96.32	96.09	94.94
DM	89.79	91.24	85.08	90.31	84.56	**91.61**	**91.61**
Avg	89.74	**94.40**	89.90	91.98	91.15	93.71	92.87

4. Higher dimension class features

When classes contain more than two class values they need to be treated as ordinal even if they are not. For example if the

three class outcomes are light, medium and heavy and we have 5 light examples, 0 medium examples and 5 heavy examples we have conditional probabilities of $Pr(light|F_v) = 0.5$, $Pr(medium|F_v) = 0.0$ and $Pr(heavy|F_v) = 0.5$. The feature value, F_v would be assigned a weight of 0 using AWSum which places it in the middle of the influence scale. In terms of conditional probability this is inconsistent as there are no medium examples, but in terms of influence on the outcome it is intuitive because we can reasonably say that the influence of 5 heavy examples and 5 light examples is the same as 10 medium examples. This approach can be demonstrated to have a good classification outcome even in cases such as the Iris dataset where the outcomes are not ordinal. However the visualisation may be misleading in that a value could appear at the middle of the scale either because there is a high probability of that outcome or because class values at the extremes have the same probability.

In order to scale the conditional probabilities that constitute the influence weight, a simple mapping value as per equation 4 is applied.

$$M_i = \left(\frac{2}{c-1} \times (i-1)\right) - 1 \tag{4}$$

where: c = the number of class values and i is the mapping value for the i^{th}class value

5. Application to the Diabetes Dataset

AWSum's ability to convey meaningful information on the influences affecting outcomes to the user has been tested using Diabetes data. This data was supplied by Charles Sturt University [12] and consists of 1930 records, 77 features, and a class with 2 values that represent a diagnosis of no diabetes and Type 2 diabetes.

In order to be useful in real world situations the insights presented need to convey meaning to the user and be easy to interpret. This was tested by giving a medical researcher, expert in the field the output from AWSum for the Diabetes data and analyzing their interpretation of the information. The second criterion measured was the accuracy of the insight. AWSum's measure of influence for single feature values and combinations of feature values was presented to the expert and his comment elicited regarding the appropriateness of the influence measure. These influence weights were presented in two different formats. The first, as seen in figures 3 and 4 show the absolute influence of the feature values without regard to the prior probability of the outcome. By this we mean that a weight of 0 for a feature value indicates that 50% of the time that value occurred the person had 'no diabetes' and 50% of the time the person had 'type2 diabetes'. The second presentation of the data adjusts the weights by the prior probability of the class value. In this case the probability in the sample of 'type 2 diabetes' is 0.26 and the probability of 'no diabetes' is 0.74. When we calculate a weight for this as we do for the features values it is -0.48. The intuition behind the second set of figures 5 and 6 is that if we knew nothing about the person in the sample their influence weight on a scale of -1 = no diabetes to 1 equal to type 2 diabetes would be -0.48. Therefore if a feature value's influence weight is less than -0.48 it could be said to increase the influence toward 'no diabetes' relative to the sample population and if it were greater than -0.48 it could be said to increase the influence toward 'type 2 diabetes' relative to the sample population. The threshold generated by AWSum for separating the two class values could be used in place of the prior probability as it will approximate it. In this case it is -0.44.

Further testing will be required on a range of datasets using a number of experts but preliminary results have been encouraging as illustrated in the following section.

A. Ease of Interpretation

The expert was presented with diagrams as described above. There were: 195 single feature values and 89 combinations of 2 feature values. For the single feature values the expert interpreted the figure as telling him that if a patient had the feature value concerned this would lead to a likelihood of diabetes as indicated by the influence weight. For the combinations of feature values the expert interpreted the combination influence weight as being the likelihood of diabetes that could be expected when these factors occurred together in a patient. The expert was able to determine that this was potentially different to the way that the constituent feature values may act when occurring independently.

These interpretations indicate that the information presented is being interpreted correctly by our expert. It needs to be noted that the expert was always eager to extrapolate causality from the influence weight. This is to be expected in a field where interventions and diagnosis are the focus.

B. Accuracy of Insights

When an insight is being assessed it falls into one of several categories:

- Correct and expected
- Correct but unexpected
- Incorrect

Insights that are correct and expected, help verify the insight process and confirm domain knowledge. Those that are unexpected need further explanation. It could be that they are incorrect, although as the weights are heavily based on conditional probabilities this would need further investigation and may imply that the data is unrepresentative of the population. The unexpected influence weights may also reflect new domain knowledge and uncover associations that may or may not be causal.

It is difficult in a field such as this to quantify exactly the level of agreement between the influence weight and the expert's domain knowledge. For this experiment the expert was simply asked to comment on the appropriateness of the influence weights presented. The expert's domain knowledge largely concurred with the influence weights presented. An exception was a high reading for "waist measurement" which the influence weight indicated was an indicator of not having diabetes but the expert felt was a clear indication of having

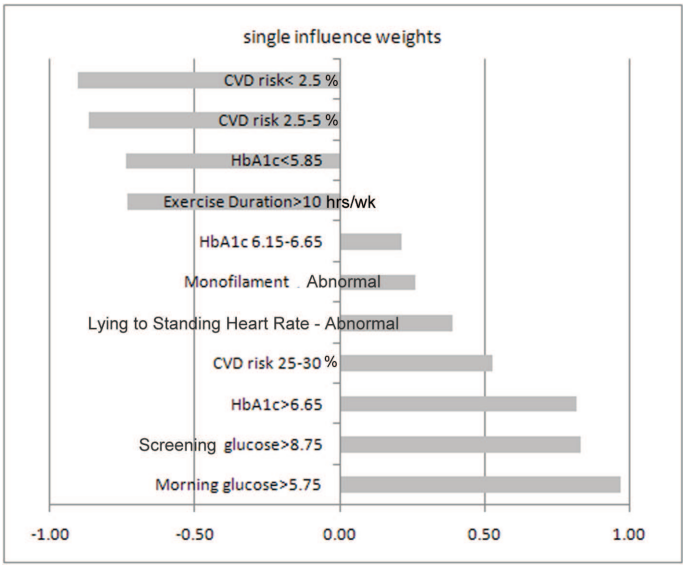

Fig. 3: Influence weights for feature values

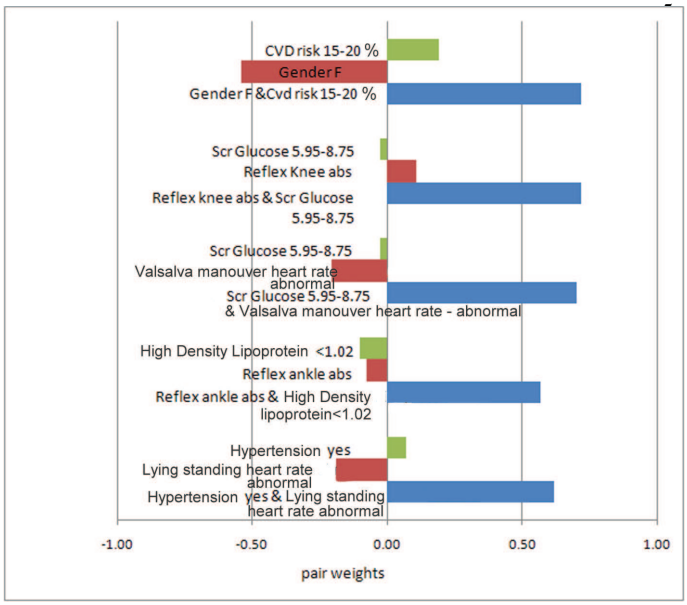

Fig. 4: Influence weights - pairs of feature values

diabetes. This difference was later identified to have been caused in the collection of the data by mixing measurement units of inches and centimetres. This is not the sort of anomaly most classifiers would identify and is a useful trait of the AWSum classifier.

Of the pairs of feature values presented the expert again largely concurred with the weights presented but interest was shown in those pairs containing an indication of an absence of reflex in the knees and ankles and a glucose reading that was high but below the diagnostic threshold for diabetes. These seemed to indicate a higher influence toward a diagnosis of diabetes than the expert would have expected. Influence weights such as these that are outside the domain knowledge of the expert can direct further research and more data is being sought in order to verify the anomalies found in the research to date.

6. Conclusion and Future Work

AWSum has demonstrated some usefulness as a potentially valuable research tool. This usefulness comes about firstly because it provides a scaled influence weight for feature values allowing comparison with the expert's own knowledge but more than that it can point to complex associations between combinations of feature values and the outcome. This becomes particularly important when this interaction is outside the expert's domain knowledge. While not implying any causality it points to an association of interest requiring explanation.

While acknowledging that further work is required before any sort of prognostic model could be proposed, due to the rigor required in the medical field, AWSum has shown that it can use historical data to identify the influence of features and combinations of features that largely concur with expert opinion. This suggests some value in pursuing an evaluation of its potential to provide decision support. The other factors pointing to this as a possible direction are the simplicity of the

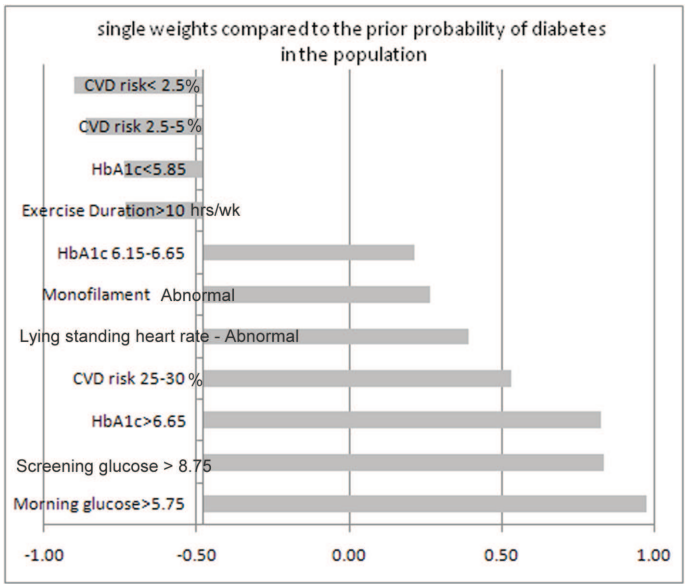

Fig. 5: Influence weights relative to prior probability of Diabetes in the sample

approach and its the understandability of its output. This can plausibly lead to the generation of diagnostic and prognostic models that are more widely adopted in clinical practice than many existing models.

References

[1] Blake, C.L., Newman, D.J., Hettich, S. and Merz, C.J.: UCI repository of machine learning databases. (1988)

[2] Duda, R., Hart, P. : Pattern Classification and scene analysis. John Wiley and Sons. (1973).

[3] Friedmann, N. and Goldszmidt: Building classifiers using Bayesian intelligence. 'In Proceedings of the National Conference on Artificial Intelligence.' AAAI Press, Portland, Oregon, USA, (1993) pp. 207–216..

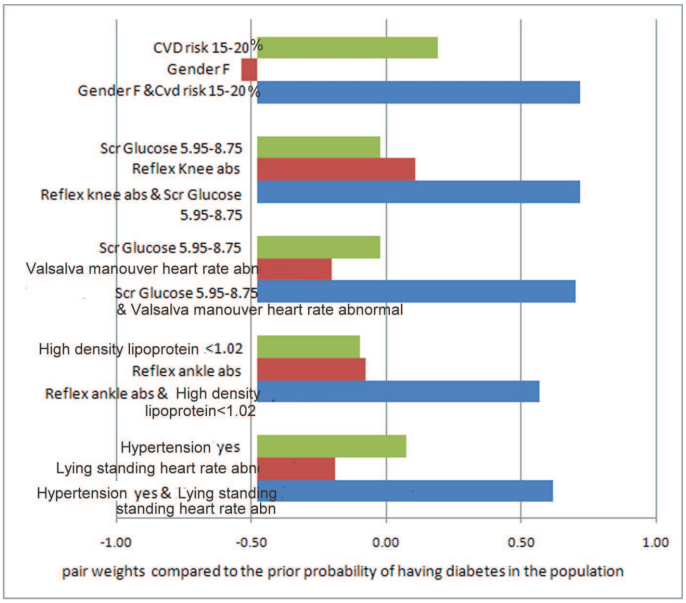

Fig. 6: Influence pairs relative to prior probability of Diabetes in the sample

[4] Quinlan, J. : Programs for Machine Learning. Morgan Kaufmann (1993).

[5] Quinn, A., Stranieri, A. and Yearwood, J. : Classification for accuracy and insight. A weighted sum approach. In proceedings of 6th Australasian data mining conference, **70** Gold Coast, Australia (2007).

[6] Setiono, R. and Liu, H. Symbolic Representation of Neural Networks, Computer, **29**,IEEE Computer Society Press, Los Alamitos, CA, USA, (1973) pp. 71–77.

[7] Shafer, G. : A Mathematical theory of evidence. Princeton University Press (1993).

[8] Teasdale C, Jennett B. Assessment of coma and impaired consciousness: a practical scale. Lancet 1974;ii:81-4.

[9] Vapnik, V.: The nature of statistical learning theory. Springer - Verlag (1999).

[10] Witten, I.H. and Frank, E.: Data Mining: Practical machine learning tools and techniques with java implementations. Morgan Kaufmann (2000).

[11] Klotz, S., Nand, K., Richard De Armond, R., Donald Sheppard D; Nancy Khardori ; John E. Edwards Jr ; Peter N. Lipkee; Mohamed El-Azizi: Candida albicans Als proteins mediate aggregation with bacteria and yeasts Medical Mycology, **45**, (2007) , pp. 363 – 370

[12] School of Community Health, Charles Sturt University; Diabetes Screening Dataset, 2002-2008

[13] Usama M. Fayyad, Keki B. Irani: Multi-interval discretization of continuous valued attributes for classification learning. In: Thirteenth International Joint Conference on Artificial Intelligence, 1022-1027, 1993.

[14] Wyatt JC, Altman DG. Prognostic models: clinically useful, or quickly forgotten ? BMJ 1995; 311: 1539-41 (commentary)