Int J Software Informatics, Vol.2, No.2, December 2008, pp. 199–214
International Journal of Software and Informatics, ISSN 1673-7288

E-mail: ijsi@iscas.ac.cn
http://www.ijsi.org
Tel: +86-10-62661040

AWSum–Combining Classification with Knowledge Aquisition*

Anthony Quinn[1], Andrew Stranieri[1], John Yearwood[1],

Gaudenz Hafen[2] , and Herbert Jelinek[3]

[1](University of Ballarat, Australia, quinn@clearmail.com.au)

[2](Department of Pediatrics, University Hospital CHUV, Lausanne, Switzerland, gaudenz.hafen@gmx.ch)

[3](Department of Community Health, Charles Sturt University, Albury-Wodonga, Australia, hjelinek@csu.edu.au)

Abstract Many classifiers achieve high levels of accuracy but have limited applicability in real world situations because they do not lead to a greater understanding or insight into the way features influence the classification. In areas such as health informatics a classifier that clearly identifies the influences on classification can be used to direct research and formulate interventions. This research investigates the practical applications of Automated Weighted Sum, (AWSum), a classifier that provides accuracy comparable to other techniques whist providing insight into the data. This is achieved by calculating a weight for each feature value that represents its influence on the class value. The merits of this approach in classification and insight are evaluated on a Cystic Fibrosis and Diabetes datasets with positive results.

Key words: data mining; classification; knowledge acquisition; weighted sum

Quinn A, Stranieri A, Yearwood J, Hafen G, Jelinek H. AWSum–Combining classification with knowledge aquisition. *Int J Software Informatics*, **2008, 2(2): 199–214. http://www.ijsi.org/1673-7288/2/199.pdf**

1 Introduction

Probably the most widely quoted definition of data mining is that of Frawley *et al.*[9]; the non-trivial extraction of implicit, previously unknown and potentially useful information from data Others have also given various definitions. Fayyad[8] describes data mining as the first step in extracting information that is understandable and informative from large volumes of raw data, while Kohavi[14] referred to the output of data mining as 'insight' which he defined as "identifying patterns and trends that are comprehensible, so that action can be taken based on the insight.".

All these definitions imply that data mining should have goals beyond accurate classification and that data mining techniques should provide insight or knowledge to the user beyond a simple classification. It has been further argued by Pazzani[21] that

* Corresponding author: Anthony Quinn, Email: quinn@clearmail.com.au
Manuscript received 15 Oct., 2008; revised 4 Dec., 2008; accepted 19 Dec., 2008; published online 26 Dec., 2008.

the knowledge produced by data mining techniques ought to be understandable and comprehensible to the user which as Clark[5] has point out in relation to the medical field is not always the case.

Kohavi[14] contends that the vast majority of research in data mining has centered on the development of predictive accuracy and that this is in part due to the fact that accuracy can be quantified whereas insight and knowledge, is harder to quantify. Insight is though necessary in many domains because the users may not be accepting of predictions coming from a source they don't fully understand or because legally they be must explain and justify predictions. This is the case in the medical profession as described by Wyatt[30].

Following Kohavi's definition, we refer to the discovery of knowledge that is comprehensible and of interest in a practical scenario as insight. The level of interest in a practical scenario is a subjective measure that alters according to the user and thus is difficult to quantify absolutely. Our approach to this has been to use subject area experts to assess the usefulness of AWSum as a tool to both confirm domain knowledge and uncover new and interesting knowledge.

This research investigates the practical applications of Automated Weighted Sum *AWSum*, using Cystic Fibrosis[2] and Diabetes[4] datasets. AWSum seeks to provide knowledge discovery in combination with predictive accuracy. The knowledge discovery component of AWSum derives from its use of associations, as would be seen in association rules, to derive a weight for each feature value that represents its scaled influence on the classification. Figure 1 demonstrates the intuition behind this approach. The process of calculating an influence weight for an association applies equally to those associations with multiple antecedents. For example an influence weight can be calculated for the influence of high blood pressure on a heart attack as can the influence weight for high blood pressure and high cholesterol on heart attack. This ability to analyse the influence of multiple factors and reduce it to a single weight has implications in practical settings where the consideration of multiple factors is required. The difficulty medical practitioners have in considering multiple factors was noted by Johnson *et al.*[12] along with the conclusion that decision support system can improve diagnosis.

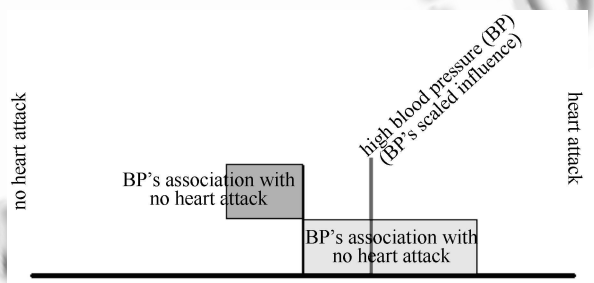

Figure 1. Combining associations to form influence weights

2 Background

The goals of extracting insight and classification are diverse, meaning that an algorithm can be very effective at one and not the other. For example association rules[1] lend themselves to insight as they inform the user of frequent item sets but they

are not though predictive unless modified for the purpose[17,15]. Neural Networks[24] on the other hand can be good at prediction but provide the user with little understanding of the process or influences on classification. AWSum seeks to address both goals.

AWSum's approach to classifying an example can be seen as a weighted sum approach or a combination of evidence. In a weighted sum approach a weight is allocated to each feature value and these weights are added together for an example and compared to a threshold in order to classify the example. This is the method adopted by AWSum although instead of the usual approach of manually assigning the weights and thresholds using domain expertise they are automatically generated from historical data. This could also be seen as a combination of evidence in that the intuition behind AWSum is that each feature value or combination of feature values has a measurable influence on classification and that these influences can be combined to indicate the class feature.

AWSum differs from most classifiers in the level of interaction the user can have with the classification process. Most classifiers run start to finish without presenting the user with any opportunity to validate or influence the measures used to classify. AWSum uses a two step approach to classification; firstly the influence weights are calculated and then they are used to classify. The expert is able to validate the influence weights at the intermediate step giving them a role and providing them with an understanding of the mechanisms being used to classify.

3 The Algorithm

The algorithm consists of 2 steps; the first involves the calculation of influence weights for each feature value and the second involves the determination of optimal threshold values for the classification of new examples.

3.1 Influence weights

The first phase of the AWSum approach lays the foundation for classification by calculating influence weights for each feature value. Calculating the conditional probability of the outcome given the feature value gives the level of association between the feature value and an outcome. To calculate an influence weight the level of association in relation to each class value, for a given feature value, is combined into a single figure.

The algorithm is described below using binary classification for simplicity. 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 values of the binary class. To achieve this the conditional probabilities associated with one class value are considered to be positive and conditional probabilities associated with 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 conditional probabilities we arrive at a single influence weight that represents the feature value's influence on one class value relative to the other. Equation 1.1 demonstrates this calculation and Fig.2 shows an

example where $Pr\,(O_1|Fv) = 0.2$, or -0.2 when mapped and $Pr\,(O_2|Fv) = 0.8$.

$$W = Pr(O_1|Fv) + Pr(O_2|Fv) \tag{1.1}$$

where W = the feature value combination influence weight
O_1 = the first outcome in a binary outcome
O_2 = the second outcome in a binary outcome
Fv = the feature value in the combination

Figure 2. Combining conditional probabilities to form an influence weight for a binary class example

3.2 Classification

Classification of an example is achieved by combining the influences 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 1.2 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.

$$e_i = \frac{1}{n}\sum_{m=1}^{n} W_m \tag{1.2}$$

where e_i = the influence weight of the i^{th} example
n = the number of features
W_m =is the m^{th} influence weight

The influence score of 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. The examples with total influence scores that fall to the left of the threshold in Figure 3 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 misclassified but the number of misclassifications has 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. 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.

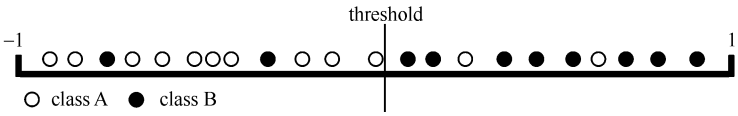

Figure 3. Classification minimisation on training example to form a threshold

An advantage provided by this optimisation approach is that a cost function is easily incorporated into the threshold calculation. This can be done by simply choosing to place the threshold at the point in the training data where there is no misclassification of a chosen class.

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[7].

3.3 Considering the influence of combinations of feature values

The combination 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 as seen in equation 1.3. These combinations of feature values can contribute to an increase in accuracy and provide insight. Analysts can 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 different this indicates a level of interaction between the feature values. This is useful, for example, in identifying things such as adverse drug reactions.

$$W = Pr\left(O_1|Fv_1, Fv_2\right) + Pr\left(O_2|Fv_1, Fv_2\right) \tag{1.3}$$

where W = the feature value combination influence weight
O_1 = the first outcome in a binary outcome
O_2 = the second outcome in a binary outcome
Fv_1 = the first feature value in the combination
Fv_2 = the second feature value in the combination

3.4 N-ary classification

In order to represent 3 or more class values on a linear scale assumptions need to be made. The class values need to be considered as ordinal. For example if the 3 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 classify well even in cases such as the Iris dataset where the outcomes are not ordinal but the visualisation may be misleading in that a value at the middle of the scale could appear there either because there is a high probability of that outcome or because class values at the extremes have the same probability. The approach was also comparable in accuracy on the Cystic Fibrosis dataset that has 3 ordinal class values. Equation 1.4 demonstrates the mapping values to be applied to each conditional probability when calculating a feature value weight for problems with two or more class values.

The equation has the effect of segmenting the -1 to 1 scale into as many equal intervals as there are class values. In a 4 class problem the mapping values would be -1, -0.33, 0.33, 1. This approach assumes a continuum exists between the classes along the -1 to 1 scale in much the same way that the sigmoid function in linear regression assumes a continuous function exists between binary outcomes when in fact the outcomes are discrete.

$$M_i = \left(\frac{2}{c-1} \times (i-1)\right) - 1 \tag{1.4}$$

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

3.5 Model selection

An influence weight for each feature value and all combinations of feature values that exist in the problem domain are calculated. In order to select which combinations of feature values to include in the classification model a comparison of the influence of the feature value combination and its parents is undertaken. 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 as seen in equation 1.5. 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 described in equation 1.2 then there is no need to include the child in the classification model. The level of significance used to identify useful combinations of feature values can be established for classification by testing for improvement in classification on the training set or in the case of identifying interesting combinations it can be arbitrarily set by the domain expert.

The ability to identify combinations of feature values that interact strongly can identify possible areas of interest for researchers.

$$W_{diff} = W_{F_1} - W_{F_1|F_2} \tag{1.5}$$

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 on the training set.

3.6 Scalability

A single pass of the dataset is sufficient to count the occurrences of each feature value and combination of feature values as they occur with each class value. The counts can then be used to calculate any required influences weights. The number of counts is constant regardless of the number of records and given the counts are performed efficiently, scalability is not compromised. Two counts are required to calculate each influence weight and so the number of influence weights to be calculated effects the scalability of the algorithm. Equation 1.6 indicates the number of influence weights calculated. It indicates that the number of influence weights is related combinatorially to the number of features and exponentially to the average number of feature values per feature.

When dealing with real world datasets such as CF and diabetes the number of influence weight is vastly reduced when compared to the potential number of combinations because many combinations do not exist in the data. We can also reduce the number of influence weights by setting a confidence and support for them in the same fashion as association rule mining. A further technique for reducing the number of influence weights is not to include those combinations that are outside an arbitrarily set threshold for difference to their constituent components influence weights.

$$S = C\,(m, r) \times n^r \tag{1.6}$$

where m = the number of features
n = the average number of feature values in each feature
r = the number of constituent feature values in each combination eg singles pairs triples etc

4 Experiments

Four datasets were sourced from the UCI Repository[3] for the comparative evaluation of the AWSum approach. In addition, the Cystic Fibrosis dataset[2], with 17 categorical features, 6 continuous features, 3 classes, 212 instances, and many missing values, and the Diabetes dataset withe 28 features, 2 classes, 1930 instances and missing values were used. Ten fold stratified cross validation was used in all experiments. Table 1 illustrates the classification accuracy by other techniques using the Weka[29] suite alongside results from AWSum. AWSum Single refers to the results using single feature 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. Table 1 illustrates that AWSum performs comparably on all datasets.

5 Application to the Cystic Fibrosis Data

AWSum's ability to convey meaningful insights to the user has been tested using Cystic Fibrosis data supplied by the ACFDR[2].

Table 1 Classifier comparison using single feature value influence weights only

Data	AWSum Single	AWSum Triple	NBC	TAN	C4.5	SVM	Logistic
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
CF	48.40	**64.24**	60.38	59.91	60.85	55.66	60.84
DM	89.79	91.24	85.08	90.31	84.56	**91.61**	**91.61**
Avg	82.85	**89.37**	84.98	86.63	86.10	87.37	87.53

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 evaluated by giving a domain expert the output from AWSum for the CF data and analyzing their interpretation of the information. The second criteria measured was the accuracy of the insight. AWSum's measure of influence for single feature values and combinations of feature values was presented to a CF expert for comments on the appropriateness of the influence measure. Preliminary results are encouraging.

5.1 Ease of interpretation

The expert was presented with diagrams in the form seen in Fig.4. There were: 21 single feature values, 25 combinations of 2 feature values and 16 combinations of 3 feature values presented. 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 level of severity of CF as indicated by the influence weight. For the combinations of feature values the expert interpreted the combination influence weight as being the level of severity 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.

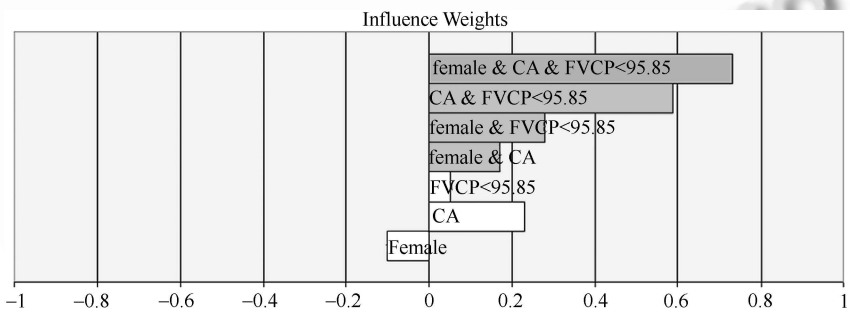

Figure 4. Influence weights for feature values and combinations of feature values

This indicates that the information presented is being interpreted correctly by our expert. It needs to be noted that the expert was always keen to interpret causality. For instance, he noted influence weights such as *presence of yeast infection candida albicans* (CA) and *breath volume* (FVCP<95.85) where he considered that the association was not causal. This is to be expected in a field where interventions and diagnosis are the focus.

5.2 Accuracy of insights

When an insight is being assessed it falls into one of several categories: *Correct and expected*, *Correct and unexpected* or *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 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 experts domain knowledge. For this experiment the expert was simply asked to comment on the appropriateness of the influence weights presented. Of the 62 influence weights the expert deemed 60 or 96.8 percent to be appropriate. It can be said that these influence weights were both correct and expected, although they do give the additional advantage of scaling and quantifying the influences, which the expert found informative and helpful.

The two influence weights that were unexpected to the expert involve the presence of Candida Albicans (CA). They were, CA and a breath volume indicator FVCP<95.85 and Female, CA and FVCP<95.85. Individually CA and FVCP<95.85 are not strong indicators of severe CF having influence weights of 0.23 and 0.05 respectively. The expert concurred with these weights. When they occur together the influence weight jumps to 0.59. This increases again to 0.73 for females with CA and FVCP<95.85. This can be seen graphically in Fig.4. In was not in the experts experience that CA had a clinical link with the severity of CF. His suggestion was that perhaps severe CF caused CA, although this explanation doesn't fully cover what is seen in the data, as CA seems to compound the CF severity when associated with FVCP<95.85. The explanation for the increase for females may be that females more often have CA. This data has proven interesting enough to the expert that further enquiries are being made of experts in the CA area to try and determine an explanation for the observation. There has also been microbiological research identified[18] that suggests a possible causal link between CA and the severity of CF. While no causal link has been established at this stage and may well not be the insight provided by AWSum has proved interesting to our expert and prompted him to consult other related experts. This indicates that AWSum can reveal insights that are complex and of interest in real world research.

6 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 collected from patients visiting a screening clinic run by Charles Sturt University[4] and consists of 1930 records, 77 features, and a class with 2 values that represent a diagnosis of no diabetes and Type 2 diabetes.

The influence weights were presented in two different formats. The first, as seen in Figs.5 and 6 shows 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 times the feature value occurred the person had 'no diabetes' and 50% of the time the person had 'type2 diabetes'.

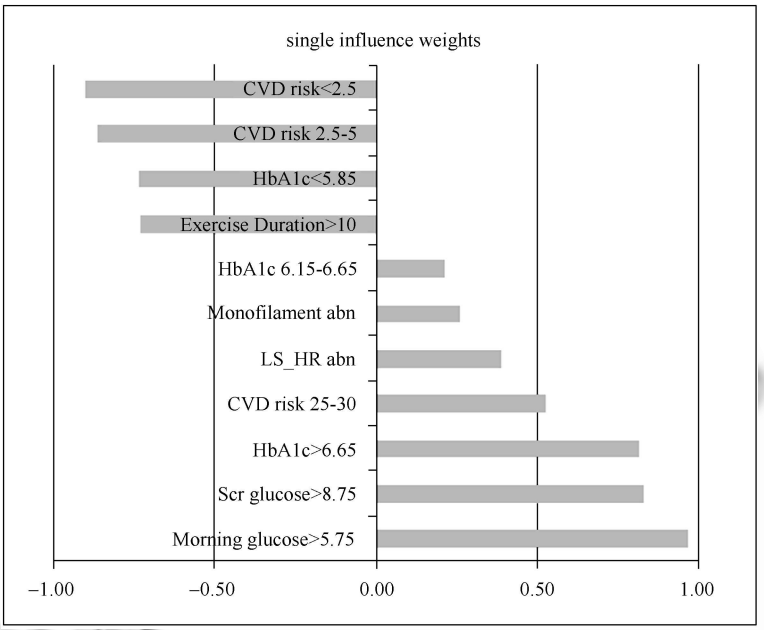

Figure 5. Influence weights for feature values

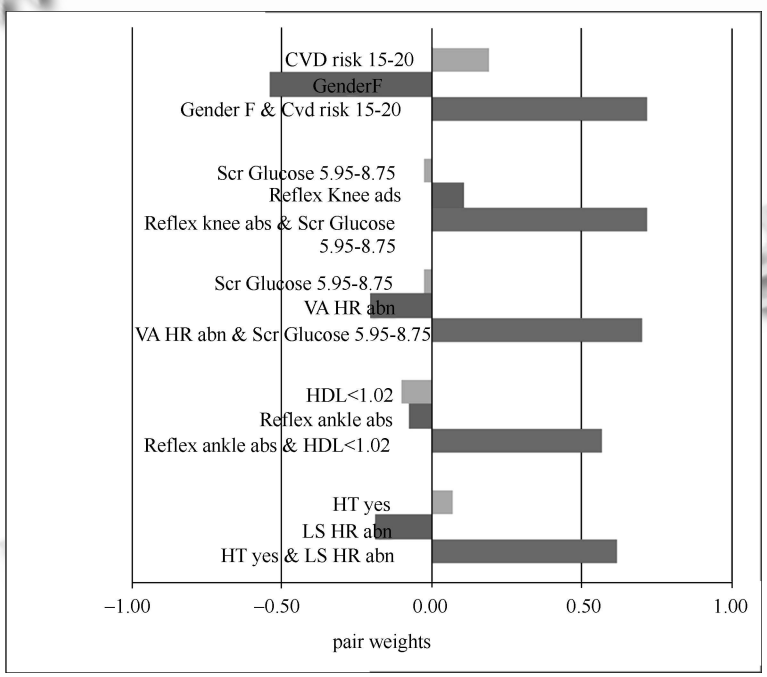

Figure 6. Influence weights-pairs of feature values

The second presentation of the data shows the influence weight relative to the prevalence of type 2 diabetes in the population as seen in Figs.7 and 8. In this case

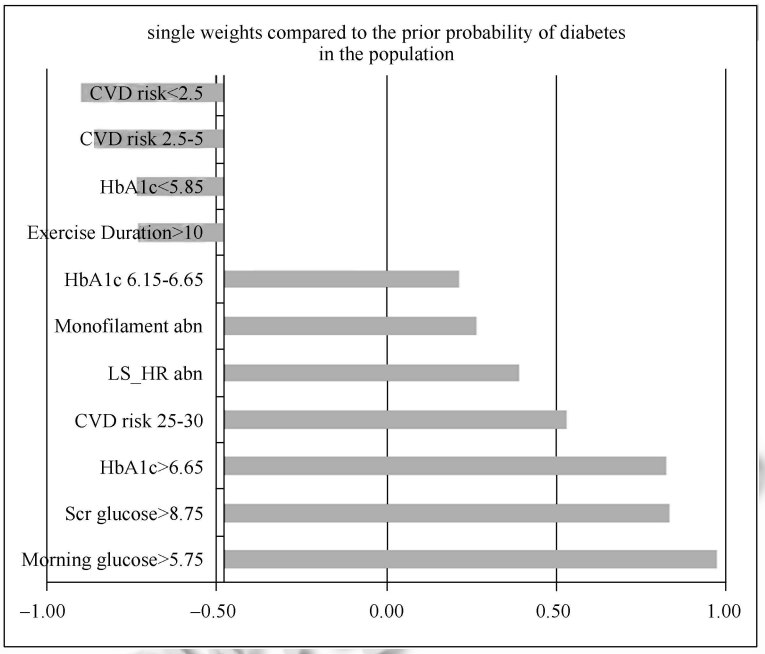

Figure 7. Influence weights relative to prior probability of Diabetes in the sample

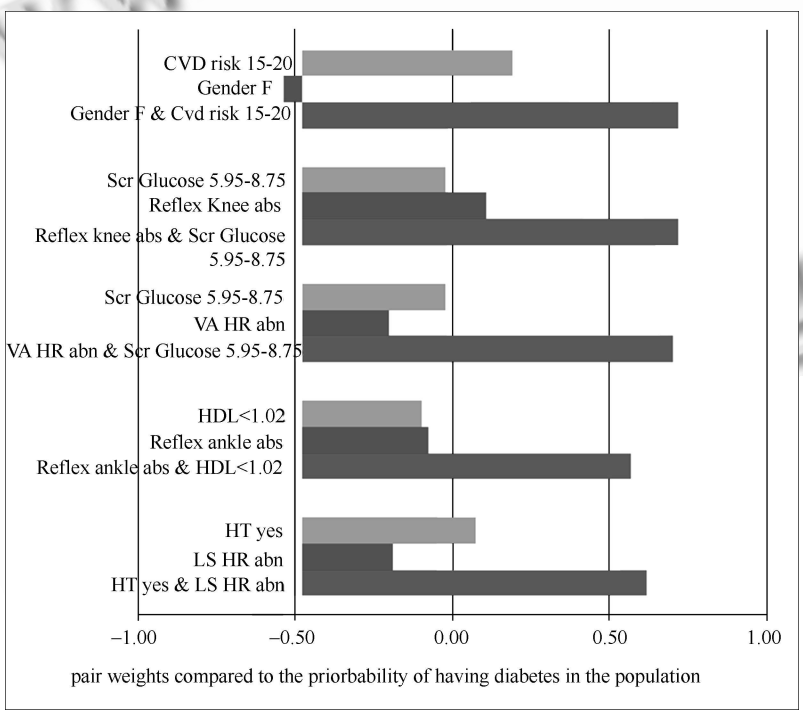

Figure 8. Influence pairs relative to prior probability of Diabetes in the sample

the probability 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 and therefore influence weights less than -0.48 increase the influence toward 'no diabetes'

relative to the sample population and those greater than -0.48 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.

6.1 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.

The form of presentation of data that proved most acceptable to the expert was as shown in Figs.7 and 8 that present influence weights relative to the prior probability of type 2 diabetes. An example of this is in the expert's interpretation of *VA* in Fig.6 as indicating an influence toward 'no diabetes' when they believed it was an influence toward 'type 2 diabetes'. This anomaly occurs because the likelihood of having type 2 diabetes due to this *VA* reading is less than 50% and therefore the influence weight is negative but at the same time the likelihood of type 2 diabetes given VA is greater than that of type 2 diabetes in the population and so it is a relative positive influence toward type 2 diabetes. This would seem to indicate that in practical medical research the intuitive approach is to view factors as having a positive or negative influence on the likelihood of disease from a baseline of the probability of the disease to begin with. This is not surprising in a field that is trying to identify causation and possible interventions.

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 diabetes. This difference was later identified to have been caused in the collection of the data by mixing measurement units of inches and centimeters. 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. Absences of reflex and high sugar levels when occurring together indicated a strong influence toward diabetes whereas their individual effects were relatively weak.. The significance of a feature pairing like this is that both reflex and sugar level are easy to measure in the field and if their combined influence were confirmed it would give clinicians an easy to obtain indication of a strong likelihood of diabetes.

The expert also noted that the gender and Cardio Vascular Disease (CVD) risk was interesting as it reflected common patho-physiological mechanisms in that the

female gender usually has less prevalence of diabetes and is protected against heart disease until menopause[20]. The increased CVD risk, which is in the high category, indicates some commonality between heart disease and diabetes progression.

This analysis by the expert was encouraging and while noting that further analysis is warranted it indicated that AWSum was capable of both confirming domain knowledge as well as identifying influences on the outcome that were of interest to the expert.

7 Discussion

An new approach such as AWSum raises many questions and the following section seeks to address some of these.

7.1 Causation

During our experimentation we noted that our subject area experts were focused on causality, as could be expected in a field such as medical research. Given that AWSum establishes influence weights based on association it is valid to question its usefulness in a field focused on causality. Research conducted by Wyatt[30] into the reasons that medical practitioners failed to take up prognostic algorithms was that they felt uncomfortable being dictated to by computerised models as to cause or diagnosis. Wyatt concluded that in order for prognostic models to be more acceptable in the field they ought seek input from the practitioner and allow their domain knowledge to be integrated into the model. For this reason we believe that AWSum's approach of suggesting an association and leaving it to the practitioner to establish cause is suitable. It should also be noted that in the medical field the any hypothesis needs to be exhaustively tested and so a causal model would not be sufficiently acceptable to the profession even if available.

7.2 Feature values verses features

AWSum focuses on the importance of feature values and combinations of feature values rather than on features when selecting a model. This is differs from the approach of most classifiers.

Probabilistic approaches such as augmented Bayes tend to select a network by using a metric to find the best candidate over the space of possible networks. one of the combinations of feature values identified as interest by our expert in CF is Female,CA and FVCP<95.85. This combination parent features do not appear as having an important association in the TAN network and so would go unnoticed.

Tree based classifiers such as c4.5 use information gain to rank the importance of features, and again the tree produced does not identify important combination. C4.5 uses information gain to select the feature that best splits the data with regard to all values of the feature and so also represents a focus at feature value level

Mathematically based approaches that look to produce a function that describes the relationship of the features to the class value such as logistic regression or neural networks are also functioning at a feature level by weighting the feature or characteristics of the feature. Statistical methods such as principal components analysis look to reduce the feature space by reducing the number of features.

It is AWSum's concentration at the feature value level and its focus on combina-

tions of feature values that are influential to classification that enables it to provide the analyst with insight into the data.

7.3 Identifying important associations

AWSum's uses of a measure of the difference between the influence weight of a combination of feature values to the individual constituent values influence weights to provide a pointer to feature values that interact in an interesting way. Equation 1.7 demonstrates this concept. This provides a method of measuring interest or important feature values that is not based on either an improvement in ability to classify nor on coverage. This allows the detection of interesting interactions that would not be seen in other techniques.

influence weight of $A \cap B$ *vs influence weight of* A *and influence weight of* B

7.4 Can a qualitative approach to assessing knowledge acquisition be justified?

Data mining has traditionally been based on proofs and measurable quantities. Kohavi[14] has identified the difficulty in measuring the amount or usefulness of information in a quantitative way as a major reason for the area of knowledge acquisition being overlooked in research. At the same time Wyatt[30]points to the importance of extracting understandable and usefully knowledge to the acceptance of data mining in practical applications.

While accepting that more experimentation needs to be undertaken to fully establish AWSum's knowledge acquisition capabilities it has shown promise in the fields it has a currently been tested on. This, we believe, is encouraging enough to continue development of the approach whilst developing methodologies for assessing knowledge discovery.

We contend that as data mining has stepped outside what might traditionally be considered acceptable to mathematics and statistics so data mining may need move to formulating qualitative methodologies for the measurement of things such as knowledge.

8 Conclusion

The application of AWSum to our chosen medical datasets has demonstrated a level usefulness for the approach simply by virtue of it having uncovered some new and interesting knowledge for our fellow researchers in the medical field. It also raises some interesting questions and issues. Whilst we were able to measure classification accuracy and compare this to other algorithms as is common practice in this field our measurement of the level insight or knowledge elicited has proven more difficult. A qualitative approach to measuring results is not usual in this field and leaves open the question of the what methodology should be used in any such assessments.

As an example what is new knowledge to a user of an algorithm with average knowledge of the subject area may just be a confirmation of domain knowledge to an expert in the area. Our research has lead experts in the Cystic Fibrosis to investigate further a link identified by AWSum that was previously unknown to them but that may have be suggested as valid in some emerging microbiological research. This being so has new knowledge been discovered or has our discovery simply pointed our experts to knowledge that they were unaware of. These arguments may seem a little

philosophical but they do point to the difficulty in measuring knowledge.

We will direct efforts in future research not only to both the practical aspects of discovering knowledge in practical applications but also to establishing an acceptable methodology for assessing the level and quality of insight and knowledge gained from data mining techniques.

References

[1] Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. In: Proc. of the Twentieth International Conference on Very Large Databases. Santiago, Chile, 1994. 487–499.

[2] Australian Cystic Fibrosis Data Registry: Cystic Fibrosis Database, 1999–2003.

[3] Blake CL, Newman DJ, Hettich S, Merz CJ. UCI repository of machine learning databases. 1988.

[4] Charles Sturt University, School of Community Health: Diabetes Screening Database, 2002-2008.

[5] Clark P, Matwin S. Using Qualitative Models to Guide Inductive Learning. Morgan Kaufmann, 1993. 49–56.

[6] Duda R, Hart P. Pattern Classification and Scene Analysis. John Wiley and Sons, 1973.

[7] Fayyad UM, Keki B, Irani KE. Multi-Interval discretization of continuous valued attributes for classification learning. In: Thirteenth International Joint Conference on Articial Intelligence. 1993. 1022–1027.

[8] Fayyad UM, Piatetsky-Shapiro G, Smyth P. Advances in knowledge discovery and data mining. From data mining to knowledge discovery: an overview. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996. 1–34.

[9] Frawley WJ, Piatetsky-Shapiro G, Matheus C. Knowledge discovery in databases: An overview. In: Piatetsky-Shapiro G, Frawley WJ, eds. Knowledge Discovery in Databases, AAAI Press/MIT Press, Cambridge, MA, 1991. 1–30.

[10] Friedmann N, Goldszmidt. Building classifiers using bayesian intelligence. In: Proc. of the National Conference on Artificial Intelligence. AAAI Press, Portland, Oregon, USA, 1993. 207–216.

[11] Hu H, Li JY. Using association rules to make rule-based classifiers robust. In: Proc. of the 16th Australasian Database Conference. Newcastle, Australia. 2005. 47–54.

[12] Johnston ME, Langton KB, Haynes BR, Mathieu A. A critical appraisal of research on the effects of computer-based decision support systems on clinician performance and patient outcomes. Annals of Internal Medicine, 1994, 120(2): 135–142.

[13] Klotz S, Nand K, De Armond R, Sheppard D, Khardori N, Edwards Jr JE, Lipkee PN, El-Azizi M. Candida albicans Als proteins mediate aggregation with bacteria and yeasts. Medical Mycology, 2007, (45): 363–370.

[14] Kohavi R. Appears in the National Academy of Engineering (NAE) US Frontiers of Engineering 2000 Data Mining and Visualization.

[15] Li W, Han J, Pei J. CMAR: Accurate and efficient classification based on multiple class-association rules. In: Proc. of the 2001 IEEE International Conference on Data Mining. IEEE Computer Society Press, 2001. 369–376.

[16] Li J, Shen H, Topor R. Mining the optimal class association rule set. Knowledge-Based System 2002, 15(7): 399–405.

[17] Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. In: Proc. of the Fourth International Conference on Knowledge Discovery and Data Mining. 1998. 27–31.

[18] Maiz M, Cuevas M, Lamas A, Sousa A, Santiago Q, Saurez S. Aspergillus fumigatus and Candia Albicans in Cystic Fibrosis: Clinical Significance and Specific Immune Response Involving Serum Immunoglobulins G, A and M. Arch Bronconeumol, 2008, 44(3): 146–151.

[19] Matthews KA, Meilahn E, Kuller LH, Kelsey SF, Caggiula AW, Wing RR. Menopause and risk factors for coronary heart disease. New England Journal of Medicine, 1989, 321(10): 641–646.

[20] Michalski R, Mozetic I, Hong J, Lavrac N. The AQ15 inductive learning system: An overview

and experiments. In: Proc. of IMAL 1986. Universite de Paris-Sud, Orsay. 1986.
[21] Pazzani MJ, Mani S, Shankle WR. Beyond concise and colorful: Learning intelligible rules. Knowledge Discovery and Data Mining. 1997. 235–238.
[22] Quinlan J. Programs for Machine Learning. Morgan Kaufmann, 1993.
[23] Quinn A, Stranieri A, Yearwood J. Classification for accuracy and insight: A weighted sum approach. In: Proc. of the 6th Austalasian Data Mining Conference. Gold Coast, Australia, 2007.
[24] Hinton GE, Rumelhart DE, Williams RJ. Learining internal representations by error back propogation. PDP Research Group, ed. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 1986, 1: 318–362.
[25] Setiono R, Liu H. Symbolic representation of neural networks. Computer, IEEE Computer Society Press, Los Alamitos, CA, USA, 1973. 71–77.
[26] Shafer G. A Mathematical theory of evidence. Princeton University Press, 1993.
[27] Vapnik V. The nature of statistical learning theory. Springer - Verlag, 1999.
[28] Warren J, Stanke J, Gadzhanova S, Misan G. General practice data mining making the best of practical and fundamental limitations. In: HIC 2003: Proc. of the Health Informatics Conference. Sydney, Australlia, 2003.
[29] Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques with java implementations. Morgan Kaufmann, 2000.
[30] Wyatt JC, Altman DG. Prognostic models: Clinically useful, or quickly forgotten? British Medical Journal, 1995, 311: 1539–41.
[31] Zhang H, Jiang LX, Su J. Augmenting naive Bayes for ranking. ICML'05: Proc. of the 22nd International Conference on Machine Learning. Bonn, Germany, ACM, New York, NY, USA, 2005. 1020–1027.