Learning Differential Diagnosis of Erythemato-Squamous Diseases using Voting Feature Intervals

H. Altay Güvenir, Gülşen Demiröz and Nilsel İlter

Addresses:

H. Altay Güvenir (**Corresponding Author**)
Bilkent University,
Department of Computer Engineering and Information Science,
06533 Ankara, Turkey
E-mail: guvenir@cs.bilkent.edu.tr
Phone: +90 (312) 266 4133
Fax: +90 (312) 266 4126

Gülşen Demiröz
Bilkent University,
Department of Computer Engineering and Information Science,
06533 Ankara, Turkey

Current address of Gülşen Demiröz:
Microsoft Corporation
Redmond, WA 98052, U.S.A.
E-mail: gulsend@microsoft.com
Phone: +1 (425) 936 6181
Fax: +1 (425) 936 7329

Nilsel İlter
Gazi University
School of Medicine
Department of Dermatology
Beşevler, 06510 Ankara, Turkey
Phone: +90 (312) 214 1080
Fax: +90 (312) 212 9018

Abstract

A new classification algorithm, called VFI5 (for *Voting Feature Intervals*), is developed and applied to problem of differential diagnosis of Erythemato-Squamous diseases. The domain contains records of patients with known diagnosis. Given a training set of such records the VFI5 classifier learns how to differentiate a new case in the domain. VFI5 represents a concept in the form of *feature intervals* on each feature dimension separately. Classification in the VFI5 algorithm is based on a real-valued voting. Each feature equally participates in the voting process and the class that receives the maximum amount of votes is declared to be the predicted class. The performance of the VFI5 classifier is evaluated empirically in terms of classification accuracy and running time.

Keywords: Machine learning, differential diagnosis, erythemato-squamous, voting feature intervals

1 Introduction

Researchers working on artificial intelligence have created many algorithms that successfully learn straightforward abilities. If the context is well-defined and the bounds of the problem can be correctly encoded for the computer, then these algorithms can often pick up a pattern and learn to predict it successfully. Inductive learning is a well-known approach to automatic knowledge acquisition of such patterns and classification knowledge from examples.

In several medical domains the inductive learning systems were actually applied; for example, two classification systems are used in localization of primary tumor, prognostics of recurrence of breast cancer, diagnosis of thyroid diseases, and rheumatology [10]. The CRLS system is a system for learning categorical decision criteria in biomedical domains [15]. The case-based BOLERO system learns both plans and goals states, with the aim of improving the performance of a rule-based system by adapting the rule-based system behavior to the most recent information available about a patient [13]. The DIAGAID is a program, using connectionist approach, to determine the diagnostic value of clinical data [7].

Classification learning algorithms are composed of two components; namely, *training* and *prediction* (*classification*). The training phase, using some induction algorithms, forms a model of the domain from the training examples encoding some previous experiences. The classification phase, on the other hand, using this model, tries to predict the class that a new instance (case) belongs to.

The main requirement for such a system is prediction accuracy. Furthermore, a classification learning algorithm is expected to have a short training and prediction time. Such a system should be robust to noisy training instances. Also, in some real-world domains, both training and test instances may have some missing values. Features (attributes) that are used to encode instances may have different levels of relevancy to the domain. A classification learning system should be able to learn and/or incorporate information about the weights of the features. Another requirement might be the comprehensibility of the learned knowledge by human experts. The advantage of this trait is two folded. First, the human experts can check and verify the learned classification knowledge before it is put to use in real-world domains. Second, some previously unknown facts and patterns may be brought to the attention of human experts, leading to interesting discoveries in the field.

Previously developed machine learning algorithms, usually, possess some of these characteristics, and fail to satisfy the others. For example, some algorithms, (e.g., nearest neighbor and instance based learning algorithms [1]) develop a model of the domain quickly, but it may take quite a long time to make a prediction using this model. On the other hand, some algorithms (e.g., neural networks) can make a fast prediction, however the knowledge they learn is hard to understand and verify by humans.

Success of a classification learning algorithm, in terms of the criteria mentioned above, is directly related to the scheme used for representing the classification knowledge learned. In this paper we present a knowledge representation technique called *voting feature intervals* (VFI, for short). Along with the learning and classification algorithms, the whole system is called VFI5. The VFI representation is based on *Feature Projections* that has been used in CFP [8] and k-NNFP [2]. The VFI5, which is a non-incremental and supervised learning algorithm, is applied to differential diagnosis of Erythemato-Squamous diseases. Here, we show that that VFI5 algorithm, using the VFI representation, results in highly accurate predictions, has short training and classification times, is robust to noisy training instances and missing feature values, can use feature weights, and produces a human readable model of the classification knowledge.

The rationale behind VFI knowledge representation is that human experts maintain knowledge in this form, especially in medical domains. The input to VFI5 training algorithm is a set of training instances that are descriptions of patients with known diagnoses. Learning from these training examples, VFI5 constructs a representation of the classification knowledge inherent in the examples. This knowledge is represented as the projections of the training dataset by *feature intervals* on each feature dimension separately. Then, for each feature dimension, projection points with similar characteristics are grouped into *intervals*. Therefor, an interval represents a set of feature values that yield the same classifications.

When diagnosing a new patient, each feature participates in the voting process and the diagnosis that receives the maximum amount of votes is predicted as the diagnosis of that patient. Since each feature participates in learning and classification independently, VFI enables an easy and natural way of handling missing feature values by simply ignoring them. That is, features whose values are unknown do not participate in the voting.

The next section will describe the VFI5 algorithm in detail. In Section 3, the problem of differential diagnosis of Erythemato-Squamous diseases is explained. Application of the VFI5 algorithm to this domain is discussed in Section 4. Section 6 describes the weights learned for the features of this domain using a Genetic Algorithm. Finally, the last section concludes with some remarks and plans for future work.

2 The VFI5 Algorithm

The VFI5 classification algorithm is an improved version of the early VFI1 algorithm [6]. Here, the VFI5 algorithm is described in detail and explained through an example.

2.1 Description of the VFI5 Algorithm

The VFI5 classification algorithm represents a concept description by a set of feature intervals. The classification of a new instance is based on a voting among the classifications made by the value of each feature separately. It is a non-incremental classification algorithm; that is, all training examples are processed at once. Each training example is represented as a vector of nominal (discrete) or linear (continuous) feature values plus a label that represents the class of the example. From the training examples, the VFI5 algorithm constructs intervals for each feature. An interval is either a *range* or *point* interval. A range interval is defined on a set of consecutive values of a given feature whereas a point interval is defined a single feature value. For point intervals, only a single value is used to define that interval. For range intervals, on the other hand, it suffices to maintain only the lower bound for the range of values, since all range intervals on a feature dimension are linearly ordered. For each interval, a single value and the votes of each class in that interval are maintained. Thus, an interval may represent several classes by storing the vote for each class.

The training process in the VFI5 algorithm is shown in Figure 1. First, the *end points* for each class c on each feature dimension f are found. End points of a given class c are the lowest and highest values on a linear feature dimension f at which some instances of class c are observed. On the other hand, end points on a nominal feature dimension f of a given class c are all distinct values of f at which some instances of class c are observed. The end points of each feature f is kept in an array $EndPoints[f]$. There are $2k$ end points for each linear feature, where k is the number of classes. Then, for linear features the list of end-points on each feature dimension is sorted. If the feature is a linear feature, then point intervals from each distinct end point and range intervals between a pair of distinct end points excluding the end points are constructed. If the feature is a nominal feature, each distinct end point constitutes a point interval.

The number of training instances in each interval is counted and the count of class c instances in interval i of feature f is represented as $interval\_count[f,\ i,\ c]$ in Figure 1. These counts for each class c in each interval i on feature dimension f are computed by the $count\_instances$ procedure. For each training example, the interval i in which the value for feature f of that training example

e (e_f) falls is searched. If interval i is a point interval and e_f is equal to the lower bound (same as the upper bound for a point interval), the count of the class of that instance (e_c) in interval i is incremented by 1. If interval i is a range interval and e_f is equal to the lower bound of i (falls on the lower bound), then the count of class e_c in both interval i and $(i-1)$ are incremented by 0.5. But if e_f falls into interval i instead of falling on the lower bound, the count of class e_c in that interval is incremented by 1 normally. There is no need to consider the upper bounds as another case, because if e_f falls on the upper bound of an interval i, then e_f is the lower bound of interval $i+1$. Since all the intervals for a nominal feature are point intervals, the effect of $count\_instances$ is to count the number of instances having a particular value for nominal feature f.

To eliminate the effect of different class distributions, the count of instances of class c in interval i of feature f is then normalized by $class\_count[c]$, which is the total number of instances of class c.

The classification in the VFI5 algorithm is given in Figure 2. The process starts by initializing the votes of each class to zero. The classification operation includes a separate preclassification step on each feature. The preclassification of feature f involves a search for the interval on feature dimension f into which e_f falls, where e_f is the value test example e for feature f. If that value is unknown (missing), that feature does not participate in the classification process. Hence, the features containing missing values are simply ignored. Ignoring the feature about which nothing is known is a very natural and plausible approach.

If the value for feature f of example e is known, the interval i into which e_f falls is found. That interval may contain training examples of several classes. The classes in an interval are represented by their votes in that interval. For each class c, feature f gives a vote equal to $interval\_vote[f,\ i,\ c]$, which is vote of class c given by interval i on feature dimension f. If e_f falls on the boundary of two range intervals, then the votes are taken from the point interval constructed at that boundary point. The individual vote of feature f for class c, $feature\_vote[f, c]$, is then normalized to have the sum of votes of feature f equal to 1. Hence, the vote of feature f is a real-valued vote less than or equal to 1. Each feature f collects its votes in an individual vote vector $\langle vote_{f,1}, \ldots, vote_{f,k} \rangle$, where $vote_{f,c}$ is the individual vote of feature f for class c and k is the number of classes. After every feature completes their preclassification process, the individual vote vectors are summed up to get a total vote vector $\langle vote_1, \ldots, vote_k \rangle$. Finally, the class with the highest vote from the total vote vector is predicted to be the class of the test instance.

With this implementation, the VFI5 algorithm is a *categorical classifier*, since it returns a unique class for a test instance [11]. A unique class is predicted for the test instance in order to compare this predicted class with the actual class of the test instance. This enables us to measure the performance of our classifiers according to the most commonly used metric, which is the the percentage of correctly classified test instances over all test instances. Instead,

$$\frac{vote[C_j]}{\sum_{i=1}^{k} vote[C_i]}$$

can be used as the probability of class C_j which makes the VFI5 algorithm a more general classifier. In that case, the VFI5 algorithm returns a predicted probability distribution over all classes. Although a class is returned as the prediction of the test instance as an output of the VFI5 algorithm , the votes received by each class is also available as an output to the user enabling him/her with the level of confidence in the prediction.

2.2 An Example

In order to describe the VFI5 algorithm, consider the sample training dataset in Figure 3. In this dataset, we have two linear features f_1 and f_2, and there are 3 examples of class A and 4 examples of class B. The intervals with their class counts constructed in the training phase of the VFI5 algorithm are shown in Figure 4. For each feature, there are 9 intervals, four of which are point intervals on end points and five of which are range intervals between end points. The lower bound of the leftmost intervals is $-\infty$ and the upper bound of the rightmost intervals is ∞. The counts of each class are shown in Figure 4 above each interval. The training process continues with computing the interval class votes determined by the relative class counts after a normalization. The normalized class votes for the constructed intervals by VFI5 are shown in Figure 5. Let us look at one interval to see how the normalized votes are computed from the class counts. The interval i_{25} on feature dimension f_2 has $interval\_count[f_2, i_{25}, \mathsf{A}] = 1$ and $interval\_count[f_2, i_{25}, \mathsf{B}] = 1$ as shown in Figure 4. The class votes are $1/3 = 0.33$ for class A and $1/4 = 0.25$ for class B. Then these votes are normalized to make the sum of votes distributed to classes equal to 1, and the normalized vote for class A is equal to $interval\_vote[f_2, i_{25}, \mathsf{A}] = 0.4$ and $interval\_vote[f_2, i_{25}, \mathsf{B}] = 0.6$ for class B.

To illustrate the classification of the VFI5 algorithm on an example, let us classify the test example $t = \langle 5, 6, ?\rangle$. This test example falls on point interval i_{16} with lower bound 5 on feature dimension f_1 and on point interval i_{26} with lower bound 6 on feature dimension f_2 shown with arrows in Figure 5. Since there are point intervals on which both $t_1 = 5$ and $t_2 = 6$ fall, the individual votes of features are taken from the corresponding point intervals.

The point interval i_{16} of feature f_1 on which $t_1 = 5$ falls votes equal to $interval\_vote[f_1, i_{16}, \mathsf{A}] =$

0 and $interval\_vote[f_1, i_{16}, \mathsf{B}] = 1$ for class A and class B, respectively. Thus, the individual vote vector of f_1 is $\mathbf{v}_1 = \langle 0, 1 \rangle$. If f_1 had been given the chance to make a prediction alone, it would have predicted class B with certainty because B has received all the vote of feature f_1 and class A has received none. On the feature dimension of f_2, the point interval i_{26} on which $t_2 = 6$ falls has a vote equal to $interval\_vote[f_1, i_{26}, \mathsf{A}] = 0.57$ for class A and a vote equal to $interval\_vote[f_1, i_{26}, \mathsf{B}] =$ 0.43 for class B. Thus, the individual vote vector of f_2 is $\mathbf{v}_2 = \langle 0.57, 0.43 \rangle$. If f_2 had been given the chance to make a prediction, it would have predicted class A. Finally, the individual votes of the two features are summed up correspondingly and total vote vector is $\mathbf{v} = \langle 0.57, 1.43 \rangle$. The VFI5 algorithm votes 0.57 for class A and 1.43 for class B, so class B with the highest vote is predicted as the class of the test example.

3 Differential Diagnosis of Erythemato-Squamous Diseases

The differential diagnosis of erythemato-squamous diseases is a difficult problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are *psoriasis*, *seboreic dermatitis*, *lichen planus*, *pityriasis rosea*, *chronic dermatitis* and *pityriasis rubra pilaris*.

These diseases are frequently seen in the outpatient departments of dermatology. At the first sight all of the diseases look very much alike with the erythema and scaling. When inspected more carefully some patients have the typical clinical features of the disease at the predilection sites (localizations of the skin where a disease preters) while another group has a typical localizations.

Patients were first evaluated clinically with 12 features. The degree of erythema and scaling, whether the borders of lesions are definite or not, the presence of itching and koebner phenomenon, the form of the papules, whether the oral mucosa, elbows, knees and the scalp are involved or not, whether there is a family history or not are important for the differential diagnosis.

For example the erythema and scaling of chronic dermatitis is less than of psoriasis, the koebner phenomenon is present only in psoriasis, lichen planus and pityriasis rosea. Itching and polygonal papules are for lichen planus and follicular papules are for pityriasis rubra pilaris. Oral mucosa is a predilection site for lichen planus while knee, elbow and scalp involvements are for psoriasis. Family history is usually present for psoriasis and pityriasis rubra pilaris usually starts during childhood.

Some patients can be diagnosed with these clinical features only, but usually a biopsy is necessary for the correct and definite diagnosis. Skin samples were taken for the evaluation of 22 histopathological features.

Another difficulty for the differential diagnosis is that a disease may show the histopathological features of another disease at the beginning stage and may have the characteristic features at the following stages. Some samples show the typical histopathological features of the disease while some do not. Melanin incontinence is a diagnostic feature for lichen planus, fibrosis of the papillary dermis is for chronic dermatitis, exocytosis may be seen in lichen planus, pityriasis rosea and seboreic dermatitis. Acanthosis and parakeratosis can be seen in all the diseases in different degrees. Clubbing of the rete ridges, thinning of the suprapapillary epidermis are diagnostic for psoriasis. Disappearance of the granular layer, vacuolization and damage of basal layer, saw-tooth appearance of retes and a band like infiltrate are diagnostic for lichen planus. Follicular horn plug and perifollicular parakeratosis are hints for pityriasis rubra pilaris.

In the dataset, the family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values. The dataset used in the experiment is summarized in Table 1.

4 Experiments

Currently, the dataset for the domain contains 366 instances. We first used all of these instances to obtain a description of the domain. The description consists of the feature intervals constructed for each feature. The intervals obtained for features f_6, f_{14}, f_{15}, f_{21} and f_{34} are shown in Figure 6.

It is clear from Figure 6 that the nonzero values of feature f_6 (polygonal papules) indicate the class C_3 (pityriasis rubra pilaris). On the other hand, the high values for f_{14} would suggest class C_1 or C_2. The feature f_{15} appears to a distinguishing feature for class C_5. However, high values of f_{21} can indicate both C_1 and C_5. Also, class C_6 appears to be a children's disease.

For a particular case, let us consider a patient, who has the following values for these features: $f_6 = 1$, $f_{14} = 0$, $f_{15} = 0$, $f_{21} = 0$, $f_{34} = 52$. For that patient, the vote vector for f_6 would be $< 0.2, 0.2, 0.01, 0.2, 0.2, 0.2 >$, for $f_{14} < 0.04, 0.05, 0.25, 0.22, 0.25, 0.19 >$, and so on. Then the votes for all classes received from all 34 features are summed. The class that receives the highest amount of votes is the class predicted.

For supervised concept learning (classification) tasks, the classification accuracy of the classifier is one measure of performance. The most commonly used metric for classification accuracy is the percentage of correctly classified test instances over all test instances. To measure the classification accuracy, 10-fold cross-validation technique is used in the experiments. That is, the whole dataset is partitioned into 10 subsets. The 9 of the subsets is used as the training set, and the tenth is used as the test set. This process is repeated 10 times once for each subset being the test set. Classification is the average of these 10 runs. This technique ensures that the training and test sets are disjoint. The VFI5 algorithm achieved 96.2% accuracy on the Dermatology dataset, which means that, out of 37, only about 1 test instance is misclassified by the VFI5 algorithm.

5 Comprehensibility of VFI5

The explanation ability of a classification process is as much important as its classification accuracy. We have shown the empirical evaluation of the VFI5 classifier in Section 4 on the Dermatology dataset. But a high prediction accuracy is not enough for a classification system; the knowledge it constructs should also be comprehensible by humans. For this purpose, we tried to visualize the concept description learned by the VFI5 classifier. Since each feature votes for each class during classification, these votes form the concept description and gives information about the relation between the values of each feature and the class label observed at that value.

The concept description learned by VFI5 for the Dermatology dataset is shown in Figure 7. For space efficiency, only a few interesting features are shown. At the top of Figure 7, a general information about the dataset is given. Then the intervals with their votes for each class are displayed, where class numbers in rectangular brackets are used for the class names of the domain (see Section 3). To the right of some of the votes a (+) or (–) or nothing meaning (o) is given, which results from the following mapping of the real-valued votes to discrete evaluations:

$$s(vote) = \begin{cases} \text{if } vote = highest \text{ and } vote - next > d, \text{then} & (+) \\ \text{else if } vote < d/2, \text{ then} & (-) \\ \text{else} & (o) \end{cases} \qquad (1)$$

where $next$ is the next highest vote after $vote$ in that interval and $d = 1/No\_Classes$. The aim of this mapping was to see the ability of the features in distinguishing between classes. When a feature value makes a class (+), it means that the instance is certainly of this class. Note that, at most one class can get a (+) evaluation. A (–) class means that the instance is certainly not of this class according to feature 1 (*erythema*) and a (o) means that this feature can not say anything about the class. Unlike (+) category, more than one class can get (–) and (o).

In Figure 7, for the first feature, four range intervals and three point intervals are shown with their votes. Near the votes their discrete mapping to either (+), (–), or nothing (meaning (o)) are also shown. When there is only a (–) evaluation without the vote shown, it means that the vote is equal to 0. Since we know that features of the Dermatology dataset take only values 0, 1, 2, or 3, there will be no instance with a feature value less than 0 and greater than 3 for instance. But since we wanted our visualization to be general, those first and last intervals are also shown to the user. When we look at the first point interval $value = 0$, we see the votes $< 0.15, 0, 0.23, 0, 0.63, 0 >$ for each corresponding class. This shows that feature 1 (*erythema*) votes more than half for class 5 (*chronic dermatitis*), nearly a quarter for class 3 (*lichen planus*), more than one-tenth for class 1 (*psoriasis*), and votes none for other classes. The zero value for feature 1 confirms class 5, rejects classes 2 (*seboreic dermatitis*), 4 (*pityriasis rosea*), and 6 (*pityriasis rubra pilaris*). These real-valued votes participate in the overall voting process as they are, since in the voting scheme of VFI classifiers, there are no thresholds.

Being the designers of these classifiers, these real-valued votes were understandable for us. But thinking of the human experts (the doctors) who collected these data for us, we thought we should transform this representation into a discrete language consisting of (+): positive, (o): neutral, (–): negative. When the value of feature 1 is equal to 0, class 5 gets a (+) in the new representation, class 2, 4 and 6 (–), and other classes (o). Note that the distinguishing labels (+) and (–) are shown whereas the (o) labels are omitted in Figure 7. This means that the value of 0 for feature 1 positively distinguishes class 5 from other classes (i.e. according to feature 1 with value zero, this patient has diagnosis 5), negatively distinguishes class 2, 4 and 6 (i.e. this patient can not have diagnosis 2, 4 and 6), and says neither "yes" nor "no" for the other classes. Not all intervals distinguish much between classes: for example when the feature 1 has a value between 1 and 3 ($1 < value < 3$), all the classes are neutral (o); that is, this range of values for feature 1 does not distinguish any class from the others. In Figure 7, the $value = 0$ point interval of feature 6 (*polygonal papules*) negatively distinguishes class 3 from the other classes all of which receive equal votes in this interval. The range interval $0 < value < 3$ and the point interval $value = 3$, on the other hand, positively distinguishes class 3 and rejects all other classes. The range interval $0 < value < 3$ plus the point interval $value = 3$ correspond to the values 1, 2 and 3, which are nonzero values, for feature 6. What VFI5 learns from the training examples is that zero value (nonexistence) of feature 6 guarantees that the patient is not of class 3 and can be of any other class. On the other hand, the nonzero value (existence) of feature 6 is a very confident positive sign for class 3 whereas a very confident negative sign for all other classes. In Figure 7, the concept learned on feature 15 *fibrosis of the papillary dermis* is also shown. Nonzero values of feature 15 significantly distinguishes class 5 (*chronic dermatitis*). Lastly, the concept description learned for feature 20 (*clubbing of the rete ridges*) conveys the information that its nonzero values positively distinguish class 1 (*psoriasis*) whereas reject all other classes.

The concept descriptions are learned by classifiers in order to be used in classification of a new instance. The performance of a classifier is measured by the ratio of the number of correctly classified test instances over total number of test instances. What is as much important as the classification accuracy is the explanation ability of the classification process. Does the classifier work like a black box or can it explain why and how it came up with the resulting classification? The VFI5 classifier can explain why and how the new instance is classified as the predicted class in terms of the individual votes of each feature given for that class. Looking at these individual votes of each feature, with what level of confidence that feature confirms (high votes) or rejects (low votes) the final prediction is obvious.

An example classification of a new instance (patient) drawn from the Dermatology domain is given in Figure 8. When these comparisons were done we have used 329 training instances to learn the concept descriptions. In Figure 8, first the feature values of the instance (properties of the patient such as the age of this patient is 34) and then the individual votes of each feature distributed among classes is shown. These votes are then summed up to get the total vote vector, from which the class with the highest vote is predicted as the class of the new instance. The VFI5 classifier predicts class 1 for this instance, which was the same as the human expert's diagnosis. This is a very confident prediction for VFI5, because the next highest vote is less than the half of the vote received by the predicted class. The individual votes for class 1 are either (+) or neutral except feature 14, moreover the (+) votes almost always appear for class 1 and there is only one (+) received by class 3 from one feature. This table of votes shown in Figure 8 is a very good explanation for the classification performed in the sense that everything is open to the user. For example, feature 20 (*clubbing of the rete ridges*) gives a vote of 1.0 for class 1 (note that votes are normalized such that the sum of votes for each class is 1.0), meaning that feature 20 says that this instance must be of class 1 and reflects its individual confirmation in the total vote. At the same time, feature 20 rejects all other classes (all other classes are (–)), meaning that this instance can not be of those classes other than class 1. Feature 34 (age) with value equal to 34 is negative for *pityriasis rubra pilaris* (class 6) and neutral for all other classes. Looking only at this feature does not say anything about the class of the instance, but still it does not reject the first class.

The classification of the VFI5 classifier may not be that much confident for all the time. Let us look at another example classification in Figure 9. The feature values, the individual votes of features, and the total votes are shown in the figure. The instance is predicted as class 2 (*seboreic dermatitis*), which is the actual class predicted by the medical expert. But the next highest vote, received by class 4 (*pityriasis rosea*), is not much different than the vote of class 2. Thus, this prediction is in fact not that much confident because the classifier chooses one class rather than the other depending on a very slight difference in the votes. If we look at the individual votes of each feature, we see that there is only one (+) from feature 4 (*itching*) for class 6 (*pityriasis rubra pilaris*); that is, no feature confidently confirms class 2 or 4. There are some (–) classes and mostly the features are neutral about the classes. When we compare the feature votes of class 2 with that of class 4, we do not find votes much different than each other except the votes of especially feature 5 (*koebner phenomenon*) and 14 (*PNL infiltrate*). Since the votes of these features support class 2 rather than class 4, they significantly affect the final prediction to be class 2. The difference between votes for these two classes is the highest in feature 14, so feature 14 with value equal to 1 seems to be the most important feature in distinguishing between class 2 and class 4. Our medical expert admitted that she also encounters the same problem of distinguishing between class 2 and class 4 as encountered by the VFI5 classifier. In this classification (Figure 9), VFI5 classified the instance correctly, but there might be instances that will be misclassified by the VFI5 classifier. However, there has been no misclassification by the VFI5 classifier among 37 instances we have tested with in our experiments.

The explanations generated by the VFI5 classifier give valuable information about the classifications such as the next possible class as well as the predicted class, the features confirming which classes and how much they confirm, the features rejecting which classes. This kind of information might help the human expert in making new classifications especially if the human expert is not experienced enough.

Although the human expert collecting the data for us is an an expert in this field, our classifier detected two of her misclassifications, that made the expert change her previous diagnosis, accepting the classification made by the VFI5 classifier. Although it is very unusual, there is a possibility that a human expert can make a mistake in classification by overlooking the value of one or more features. However, a well trained mechanical classifier will consider all the features.

In this section, we have shown that the VFI5 classifier does not work like a black box and can explain why and how it came up with the resulting classification in a comprehensible way to human. The human expert agrees with the information visualized in the concept descriptions learned by the VFI5 classifier. The classification explanations do not display only the prediction but also tells how certain that prediction is compared to other classes.

6 Learning Feature Weights using a Genetic Algorithm

In a real-world domain, just like the one used in this paper, all the features used in the descriptions of instances may have different levels of relevancy. Therefore, many feature selection and feature weight learning algorithms have been developed by machine learning researchers[3, 12].

We had developed a genetic algorithm for learning the feature weights to be used with the Nearest Neighbor classification algorithm. We applied the same genetic algorithm to determine the weights of the features in our domain to be used with the VFI5 algorithm.

The weights of the 34 features, as determined by the genetic algorithm, are shown in Table 2. According to the table, koebner phenomenon has the highest weight 0.0620. Inflammatory mononuclear infiltrate is also important in the classification, with the weight of 0.0527. On the other hand, the features acanthosis, follicular horn plug, munro microabcess, and age are found to be the least relevant.

In order to assess the impact of the feature weights learned by the genetic algorithm, we have repeated the same 10-fold cross-validation experiment incorporating these weights. Using these weights, the VFI5 algorithm has achieved 99.2% accuracy, which means almost perfect classification.

7 Conclusions

In this paper, a new classification algorithm called VFI5 is developed and applied to differential diagnosis of Erythemato-Squamous diseases. Since each feature is processed separately, the missing feature values that may appear both in the training and test instances are simply ignored in VFI5. In other classification algorithms, such as decision tree inductive learning algorithms, the missing values require extra care [14]. This problem has been overcome by simply omitting the feature with the missing value in the voting process of VFI5. Also note that the VFI5 algorithm, in particular, is applicable to concepts where each feature, independent of other features, can be used in the classification of the concept. One might think that this requirement may limit the applicability of the VFI5, since in some domains the features might be dependent on each other. Holte has pointed out that the most datasets in the UCI repository are such that, for classification, their attributes can be considered independently of each other [9]. Also Kononenko claimed that in the data used by human experts there are no strong dependencies between features because features are properly defined [10]. Another advantage of the VFI5 classifier is that instead of a categorical classification, a more general probabilistic classification where the classifier returns a probability distribution over all classes is possible to implement with VFI5.

The genetic algorithm that we developed for learning relative feature weights determined the weights of the features in our domain of differential diagnosis of Erythemato-Squamous diseases. With these weight settings the VFI5 algorithm has achieved almost perfect classification. As a future work, we plan to learn and associate weights to intervals, since pure intervals representing only a single class might be more effective in classification.

Acknowledgments

This project is supported by TUBITAK (Scientific and Technical Research Council of Turkey) under Grant EEEAG-153. The authors thank Narin Emeksiz for preparing the user interface for the VFI5 program.

References

[1] D. W. Aha, D. Kibler and M. K. Albert, Instance-Based Learning Algorithms, *Machine Learning 6*, (1991) 37-66.

[2] A. Akkuş, and H. A. Güvenir, K Nearest Neighbor Classification on Feature Projections, in: *Proceedings of ICML'96* (1995) 12-19.

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

[4] S. Cost and S. Salzberg, A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features, *Machine Learning 10* (1993) 57-78.

[5] G. Demiröz and H. A. Güvenir, Genetic Algorithms to Learn Feature Weights for the Nearest Neighbor Algorithm, in: *Proceedings of BENELEARN-96* (1996) 117-126.

[6] G. Demiröz and H. A. Güvenir, Classification by Voting Feature Intervals, in: *Proceedings of Ninth European Conference on Machine Learning*(*ECML-97*), (Springer-Verlag, LNAI 1224, 1997) 85-92.

[7] J. Forsström, P. Eklund, H. Virtanen, J. Waxlax and J. Lähdevirta, DIAGAID: a connectionist approach to determine the diagnostic value of clinical data. *Artificial Intelligence in Medicine 3*, (1991), 193-201.

[8] H. A. Güvenir and İ. Şirin, Classification by Feature Partitioning. *Machine Learning 23* (1996) 47-67.

[9] R. C. Holte, Very simple classification rules perform well on most commonly used datasets. *Machine Learning 11* (1993) 63-91.

[10] I. Kononenko, Inductive and Bayesian Learning in Medical Diagnosis. *Applied Artificial Intelligence 7* (1993) 317-337.

[11] I. Kononenko and I. Bratko, Information-Based Evaluation Criterion for Classifier's Performance. *Machine Learning 6* (1991) 67-80.

[12] H. Liu and R. Setino, A probabilistic approach to feature selection - A filter solution, in: *13th International Conference on Machine Learning (ICML'96)* (1996) 319-327.

[13] B. Lopez and E. Plaza, Case-based learning of plans and goal states in medical diagnosis, *Artificial Intelligence in Medicine 6* (1997) 29-60.

[14] J. R. Quinlan, Unknown attribute values in induction. in: *Proceedings of the 6th International Workshop on Machine Learning* (1989) 164-168.

[15] A. K. Spackman, Learning Categorical Decision Criteria in Biomedical Domains, in: *Proceedings of the Fifth International Conference on Machine Learning*, (University of Michigan, Ann Arbor, 1988) 36-46.

Figure and Table Captions

Figure 1. Training phase in the VFI5 Algorithm.

Figure 2. Classification in the VFI5 Algorithm.

Figure 3. A sample training dataset with two features and two classes.

Figure 4. Intervals constructed by VFI5, along with their class counts for the sample dataset.

Figure 5. Classification of a new instance $\langle 5,\ 6,\ ?\rangle$ using the intervals constructed.

Figure 6. Some features and their intervals. The amount of vote for each class is given above the intervals. The classes that an interval gives 0 vote are not displayed.

Figure 7. Concept Description Learned by VFI5 including only a few features.

Figure 8. A correct classification of a given test instance (patient) drawn from the Dermatology domain by the VFI5 classifier.

Figure 9. Another correct (not that confident as the previous classification) classification of a given test instance (patient) drawn from the Dermatology domain by the VFI5 classifier.

Table 1. The dataset used in the experiments.

Table 2. Weights of the features as determined by the genetic algorithm.