Annals of University of Craiova, Math. Comp. Sci. Ser.
Volume 36(2), 2009, Pages 123–130
ISSN: 1223-6934

Evaluation of the Feature Space of an Erythematosquamous Dataset Using Rough Sets

Kenneth Revett, Florin Gorunescu, Abdel-Badeeh Salem,
and El-Sayed El-Dahshan

Abstract. The differential diagnosis of erythematosquamous diseases remains a difficult task requiring both clinical and histopathological data to support a diagnosis. The principle reason for diagnostic ambiguity is based on the significant degree of overlap in the overt symptoms of this class of disease. Histopathological evidence can assist in making a positive diagnosis - but is labor and resource intensive. In order to evaluate the diagnostic veracity of clinical versus histopathological features of erythematosquamous diseases, a comparison of both features classes was evaluated using rough sets. The results indicate that the histopathological feature space provided a much more significant classification rate relative to clinical features. In addition, the results of this preliminary study indicate that only a small subset of the histopathological feature space is required for maximal classification accuracy.

Key words and phrases. Datamining, Dermatology, Erythematosquamous Diseases, Reducts, Rough Sets.

1. Introduction

Dermatology is the study of disorders/diseases of the skin, hair, and nails. One particular class of dermatological disorders is termed erythematosquamous - which indicates produces redness of the skin (erythema) caused by the loss of skin cells (squamous). Typically, these disorders/diseases have either a genetic or environmental cause - and tend to occur at specific periods in life (i.e late childhood/early adolescence). The distinction between disease and disorder is typically one of degree - but generally it is held that disorders are not typically life threatening. Given that most dermatological conditions are not life-threatening (and hence typically referred to as disorders), they are never the less important to investigate since they have a significant prevalence throughout the world. Exact figures regarding incidence/prevalence is difficult to obtain - as there are many instances that occur as a result of changes in life-style. In a study published by the Riyadh Military Hospital, life-style changes such as dietary habits (e.g. consuming excess levels of chocolates) can cause dermatological disorders, which disappear when the behavior pattern changes [1].

Erythematosquamous disorders is a generic category for a variety of medical conditions that present with overlapping symptoms. There are three primary methods for diagnosing dermatological conditions: clinical, histopathological, and recently specialised microscopic examination have provided a new approach [2]. Typically, the clinical approach relies on non-invasive examination of the symptoms - such as location, size, presence of pustules, color, and related features. Histopathological examination requires extracting a physical tissue sample (biopsy) , and possibly blood work for the detection of potential viral sources for the disorder. recently, a non-invasive approach based on con-focal microscopy has been employed with success [2]. This approach provides a more robust scientific approach than pure clinical observations, without the invasiveness of the histopathological approach. The principle goal of this study was to evaluate the relative diagnostic potential of clinical versus histopathological features in a dataset containing data on 365 patients diagnosed with a variety erythemata-squamous diseases.

1.1. Previous work. The literature examining the feature space of erythematosquamous diseases is relatively recent - most studies have deployed the UCI housed dermatology dataset which was employed in this study. Guvenir and colleagues published several studies describing the use of their new classification scheme based on *voting feature intervals* (VFI) on this dataset [3,4]. The results of this study indiacte that the dataset could be classified correctly with a high degree of accuracy (approximately 99%). The authors discussed the differences between the clinical and histopathological features - but their classification analysis did not make such a distinction - instead the authors used the complete dataset for classification purposes. The VFI (and also an enhanced version termed VFI5) produces a rule set - but the rules are based on intervals and described by numeric values for the interval fiduciaries - making them less than directly interpretable by the uninitiated. Castellano and colleagues deployed a multistep neuro-fuzzy rule based system (KERNEL) to classify this dataset [5]. The KERNEL system employs a three stage approach - first a supervised clustering algorithm is deployed to extract information directly from the dataset. Next, a supervised learning based fuzzy-neural network was deployed to tune the parameters of the fuzzy rule base. Lastly, a rule refinement process is applied to produce the final minimalised rule set. The authors indicate that the age feature reduced the classification accuracy of the dataset when deploying their KERNEL method (47-71% correct classification accuracy). The authors indicate that removal of this feature significantly enhanced the classification accuracy - with classification accuracy on the order of 95%. [5] Several authors have employed a support vector machine (SVM) approach to classifying this dataset [6],[7]. In a report by Nanni, an ensemble of classifiers, each utilising a subset of the data - are combined to form a holistic classification scheme that is superior to vanilla SVM for this particular dataset. The author reports an error rate between 2-3% (as low as 0.8% in some instances) for this particular dataset using their system. Derya & Dogdu published the results of a study where they deployed a k-Means clustering approach to classifying this dataset [8]. The results from this study indicate approximately 94% overall classification accuracy when using 5 out of the 6 decision classes (omitted Pityriasis rubra pilaris - 20 instances). Czibula and colleagues discuss the deployment of a multi-agent decision support system to classifying the dermatology dataset employed in this study [9]. The authors present a generic architecture termed the multiagent decision support system (MDSS) that deploys a neural network (trained using backpropagation) to perform the classification task. Although the specific implementation details of this system are lacking in this paper, the results indicate an impressive classification accuracy of over 99%. Karabatak & Ince present an interesting approach to classification based on a hybrid approach employing association rules as input to a neural network[10]. Association rules reflect useful/interesting information contained within a database (see [11]-[13] for more details on association rules). Association rules (AR) by definition reduce the feature space of a dataset by extracting useful relations that preserve the information content (the authors in this study deployed the apriori algorithm for AR mining). This reduced feature set is then fed to a feedforward neural network, trained using the Levenberg-Marquardt back-propagation algorithm. The AR reduced the feature space from 34 to 20 attributes, and the hybrid approach produced classification accuracy on the order of 99%. A variety of neural network and fuzzy logic based approaches have been applied to this dataset, with significant classification rates (see [14],[15] for details).

Although not exhaustive, this survey hopefully provides some indication of the different approaches used to investigate the information content of this dataset. What is immediately evident is the high classification accuracy of the various approaches - many approaching 99%+. This is interesting considering the dataset contains six decision classes - with a relatively small number of records (366). This may be the result of the large feature space (34). The fundamental research question addressed in this paper is to investigate the importance of the feature space. There are two basic sets of features clinical and histopathological - with a cardinality of 12 and 22 features respectively. The first task in analysing the feature space would seem to be to determine the information content of each subset of the feature space. In order to achieve this goal, we have deployed the use of rough sets, which produces decision rules (similar but distinct from association rules) that are used to classify objects within a dataset in a supervised fashion. The rough sets approach to data classification is presented in the next section.

1.2. Rough sets. Rough sets is a relatively new data-mining technique used in the discovery of patterns within data first formally introduced by Pawlak in 1982 [16] see also [17] for a general discussion on the topic). Since its inception, the rough sets approach has been successfully applied to deal with vague or imprecise concepts, extract knowledge from data, and to reason about knowledge derived from the data. The basic philosophy of rough sets is to reduce the elements (attributes) in a DT based on the information content of each attribute or collection of attributes (objects) such that the there is a mapping between similar objects and a corresponding decision class. In general, not all of the information contained in a DT is required: many of the attributes may be redundant in the sense that they do not directly influence which decision class a particular object belongs to. One of the primary goals of rough sets is to eliminate attributes that are redundant. Rough sets use the notion of the lower and upper approximation of sets in order to generate decision boundaries that are employed to classify objects. Consider a decision table A = (U, A $\subset$ d) and let X $\subset$ U. What we wish to do is to approximate X by the information contained in B by constructing the B-lower (BL) and B-upper (BU) approximation of X. The objects in BL (BLX) can be classified with certainty as members of X, while objects in BU are not guaranteed to be members of X. The difference between the 2 approximations: BU - BL, determines whether the set is rough or not: if it is empty, the set is crisp otherwise it is a rough set. What we wish to do then is to partition the objects in the DT such that objects that are similar to one another (by virtue of their attribute values) are treated as a single entity. The rough set theory has been proved to be very useful in practice as clear from the record of many life applications; e.g. in medicine, pharmacology, engineering, banking, financial and market analysis 9see [18] for a biomedical domain example). This theory provides a powerful foundation to reveal and discover important structures in data and to classify complex objects. One of the main advantages of rough set theory is it does not need any preliminary or additional information about data. For more details on this approach to data mining, software and applications of rough sets we refer the reader to the monograph by Pawlak [17].

1.3. The dataset. The dataset used in this study (obtained from the UCI dataset repository)contains the records of 366 patients that were diagnosed with a variety of erythematosquamous diseases (the structure of the dataset is presented in Table 1). The diseases examined in this study are: psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra pilaris. For each patient, there are 34 attributes and a decision class indicating which disease the patient was diagnosed with. The values for each attribute (except for age) are nominal, with values between 0 and 3. A value of 0 indicates the lack of the feature, a value of 3 indicates a maximal value for the presence of the feature, and a value of 1 or 2 indicate relative intermediate values. The age attribute consists of actual numeric values, with a range of 17-69. There are six decision classes (diagnoses), the distribution of cases for each class is presented in table 2. There are 8 missing values in this dataset - all for the age feature. As discussed in the results section - missing values can be handled by either eliminating the record(s) or performing imputation. In this study, the dataset was used in a supervised manner, deploying the rough sets approach to classification. As discussed in the next section, rough sets is a rule based approach to classification - which was used to examine the clinical features alone, the histopathological features in isolation, and finally their combination. The purpose of this study was to determine which set of features provides the most accurate basis set for the differential diagnosis of erythematosquamous diseases.

2. Results

The first experiment conducted with this dataset was to classify all six classes using the Rosetta (v 1.41) version of an implementation of rough sets. Note that the age attribute is linear as opposed to a range of values representing degree (0-3), and in addition, several authors have noted that the Age attribute/feature reduced the classification accuracy. The results of the rough sets approach can be summarised in a confusion matrix. In table 3,a confusion matrix is presented indicating the classification results for the full dataset, without the Age attribute. The next stage was to evaluate the information content of the clinical and histopathological set of features. The dataset was split into clinical and histopathological (containing 11 and 22 features respectively), and the Age feature was excluded from the clinical feature subset. The confusion matrices for these two dataset subsets are presented in Tables 4 & 5 respectively. lastly, a random sample of rules was extracted from the rule set and is presented in Table 6.

3. Conclusion

The results from this preliminary analysis of the dermatology dataset from the UCI indicate was analysed for the information content of the feature set. There were a total of 34 features - and six decision classes in this dataset. The features were assigned relative values on a scale of 0-3, except for the family history feature (binary - yes/no), and the Age feature, which was a linear value. From table 2, it is clear that there approximately 50%of the feature set yielded correlation large coefficients with respect to their decision classes (17 features with coefficients 30% or more). It is also noteworthy that the majority of the features (28 out of 33, excluding the 'Age' feature) contained '0' as the majority value for the feature. It is interesting to note that the lack of a feature (indiacted with a '0') was so prevalent in this dataset - and may indicate that only those with a non-zero value should be considered? In addition, there were no features that were strongly correlated with the maximal value

Table 1. The decision table attributes and their correlation coefficients, along with the value for the feature that occurred most frequently. Note that the age feature is not listed. In the last column, the parenthetical value is the percentage that feature value was recorded.

Attribute name	Correlation Coefficient	Value
erythema	-0.34	2 (59%)
scaling	-0.47	2(53%)
definite borders	-0.39	2(46%)
itching	0.05	0(32%)
koebner phenomenon	-0.09	0(61%)
polygonal papules	0.06	0(81%)
follicular papules	0.48	0(91%)
oral mucosal involvement	0.06	0(82%)
knee and elbow involvement	-0.38	0(69%)
scalp involvement	-0.53	0(72%)
family history, (0 or 1)	-0.14	0(87%)
melanin incontinence	0.06	0(81%)
eosinophils in the infiltrate	-0.06	0(89%)
PNL infiltrate	-0.55	0(64%)
fibrosis of the papillary dermis	0.53	0(85%)
exocytosis	0.28	0(32%), 2(35%)
acanthosis	-0.08	2(57%)
hyperkeratosis	-0.05	0(62%)
parakeratosis	-0.42	2(36%), 1(32%)
clubbing of the rete ridges	-0.67	0(69%)
elongation of the rete ridges	-0.36	0(54%)
thinning of the suprapapillary epidermis	-0.68	0(70%)
spongiform pustule	-0.45	0(81%)
munro microabcess	-0.52	0(78%)
focal hypergranulosis	0.06	0(81%)
disappearance of the granular layer	-0.43	0(75%)
vacuolisation and damage of basal layer	0.06	0(80%)
spongiosis	0.21	0(54%)
saw-tooth appearance of retes	0.06	0(80%)
follicular horn plug	0.43	0(94%)
perifollicular parakeratosis	0.46	0(94%)
inflammatory monoluclear inflitrate	-0.02	2(56%)
band-like infiltrate	0.06	0(79%)

for any feature (indicated by a value of '3'). These observations indicate that the intermediate values formed the basis of this decision table. Internediate values are somewhat less precise than the extreme values in many cases - and may account for the large number of decision rules that were generated from this dataset (typically over 10,000), each with small support in many cases. The classification accuracy produced by the rough sets approach was reasonably high, with values exceeding 98% in some instances, depending on the set of features employed. The results from this study

TABLE 2. The labels for the six decision classes and the number of exemplars for each. Note the number in the first column is the numeric label used in the confusion matrices presented in this work.

decision class label	Number of instances
1 psoriasis	112
2 seboreic dermatitis	61
3 lichen planus	72
4 pityriasis rosea	49
5 cronic dermatitis	52
6 pityriasis rubra pilaris	20

TABLE 3. Confusion matrix (selected randomly) for the classification using the full dataset *without* the age feature. Note the overall accuracy is placed at the lower right hand corner of the table.

	1	2	3	4	5	6	
1	44	0	0	0	1	0	0.98
2	0	17	0	1	0	0	0.94
3	1	0	18	0	0	0	0.95
4	0	1	0	12	0	0	0.92
5	1	0	0	0	9	0	0.9
6	0	0	0	0	0	5	1
	0.98	0.95	1	0.92	0.9	1	*0.95*

TABLE 4. Confusion matrix (selected randomly) for the classification using the clinical featureset *without* the age feature. Note the overall accuracy is placed at the lower right hand corner of the table.

	1	2	3	4	5	6	
1	41	0	1	1	2	0	0.91
2	2	15	0	1	0	0	0.83
3	1	2	14	1	1	0	0.74
4	0	1	0	10	2	0	0.77
5	1	0	1	0	8	0	0.8
6	0	1	1	0	0	3	0.6
	0.91	0.83	0.74	0.77	0.8	0.6	*0.76*

indicate that the clinical features (with and without 'Age') produced classification accuracies on the order of 50-60%, though the results were consistently lower when the 'Age' feature was used. The histopathological features alone produce considerably higher classification accuracies of approximately 85% (again, lower when the 'Age' feature was used). The full dataset provided the largest classification accuracy (again without the 'Age' feature), on par with other published results. The rules contained approximately 10 features (average length of the left hand side), indicating that the majority of the features were superfluous. In order to investigate this result further, this dataset should be examined more closely with respect to the assignment of feature values. A more quantitative recording of feature values - and/or a different recording criteria may be required to capture more relevant information. This requires close

Table 5. Confusion matrix (selected randomly) for the classification using the histopathological featureset *without* the age feature. Note the overall accuracy is placed at the lower right hand corner of the table.

	1	2	3	4	5	6	
1	*41*	0	2	0	2	0	0.91
2	0	*17*	0	1	1	1	0.84
3	1	1	*15*	0	0	1	0.72
4	0	1	1	*11*	0	0	0.85
5	1	0	0	0	*9*	0	0.9
6	1	0	0	0	0	*4*	0.8
	0.91	0.79	0.83	0.92	0.75	0.8	*0.85*

Table 6. A sample of rules randomly selected from the histopathological feature set without the 'Age' feature. Note that a (0) and (1) indicate absence and presence of that feature respectively.

melanin incont(0) AND Eosino infiltrate(0) AND PNL infiltrat(2) AND acanthosis(1) AND hyperkeratosis(2) AND parakeratosis(1) AND munro microabcess(1) AND focal hypergranulosi(0) AND disappear granular layer(2) AND vacuolisation/damage basal layer(0) AND spongiosis(0) AND follicular horn plug(0) AND inflamm monoluclear inflitrate(2) AND band-like infiltrate(0) == Decision(1)
Eosino infiltrate(0) AND PNL infiltrat(0) AND Fibrosis papillary dermis(0) AND exocytosis(3) AND acanthosis(2) AND hyperkeratosis(2) AND parakeratosis(1) AND elong rete ridges(0) AND focal hypergranulosi(0) AND disappear granular layer(1) AND vacuolisation/damage basal layer(0) AND spongiosis(2) AND saw-tooth retes(0) AND inflamm monoluclear inflitrate(2) == Decision(4)
PNL infiltrat(1) AND Fibrosis papillary dermis(0) AND acanthosis(1) AND hyperkeratosis(0) AND parakeratosis(1) AND elong rete ridges(0) AND focal hypergranulosi(0) AND disappear granular layer(1) AND spongiosis(3) AND saw-tooth retes(0) AND follicular horn plug(0) AND inflamm monoluclear inflitrate(2) AND band-like infiltrate(0) == Decision(4)

collaboration between domain experts (e.g. dermatologists) and datamining experts. This is an area for future work.

4. Acknowledgements

The author would like to acknowledge the source of the dataset used in this study: H. Altay Guvenir, Bilkent University, Department of Computer Engineering and Information Science, 06533 Ankara, Turkey (at the time of submission) http://archive.ics.uci.edu/ml/datasets/Dermatology

References

[1] Al-Zoman, A.Y. & Al-Asmari, A.K., Pattern of skin diseases at Riyadh Military Hospital, Egyptian Dermatology Online Journal 42), December (2008)

[2] Koller, S., Gerger, A., Ahlgrimm-Siess, V., Weger, W., Smolle, J., & Hofmann-Wellenhof, R., In vivo reflectance confocal microscopy of erythematosquamous skin diseases, Experimental Dermatology. 18(6):536–540, (2009)

[3] Karabatak, M. & Ince, M.C., A new feature selection method based on association rules for diagnosis of erythemato-squamous dieases, Expert Systems with Application, pp. (2009)

[4] Guvenir, H.A., Demiroz, G., & Ilter, N., Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals, Artifial Intelligence 13, pp. 147–165 (1998)

[5] Guvenir, H.A. & Emeksiz, N., An expert system for the differential diagnosis of erythemato-squamous diseases, Epert Systems with Applications, 18(1), pp. 39–43, (2000)

[6] Castellano, G., Castiello, C., Fanelli, A.M., & Leone, C., Diagnosis of dermatological diseases by a neuro-fuzzy system, Expert System with Applications 18, pp.43–49 (2000)

[7] Nanni, L., An ensemble of classifiers for the diagnosis of erythemato-squamous diseases, Neurocomputing 69, pp. 842–845 (2006)

[8] Ubeyli, E.D., Multiclass support vector machines for diagnosis of erythemato-squamous diseases, Expert Systems with Application, 35, pp. 1733–1740, (2008)

[9] Ubeyli, E.D. & Dogdu, E., Automatic Detection of Erythemato-Squamous Disease using *k*-Means Clustering, J Med Syst, (2008)

[10] Czibula, G.S., Guran, A.M., Cojocar, G.S., & Czibula, I.G., Multiagent decision support systems based on supervised learning,

[11] Karabatak, M. & Ince, M.C., A new feature selection method based on association rules for diagnosis of erythemato-squamous dieases, Expert Systems with Application, pp. (2009)

[12] Koh Y., Finding sporadic rules in the diagnosis of the Erythemato-Squamous diseases, Intelligent Data Analysis 12(6), 621-637, (2008)

[13] Agrawal, R., Imielinski, T., & Swami, A., Mining Association Rules Between Sets of Items in Large Databases, SIGMOD Conference 1993, 207–216

[14] Ubeyli, E.D., Combined neural networks for diagnosis of erythemato-squamous diseases, Expert Systems with Applications,36(3), Part 1, pp. 5107–5112, (2009)

[15] Ubeyl,E. & Guler,I., Automatic detection of erythemato-squamous diseases using adaptive neuro-fuzzy inference systems Computers in Biology and Medicine, 35(5), pp. 421–433, (2008)

[16] Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A.: Rough sets: A tutorial. In: S.K. Pal, A. Skowron (eds): Rough Fuzzy Hybridization A New Trend in Decision Making. Springer Verlag, pp. 3-98 (1999)

[17] Pawlak, Z.,Rough Sets: Theoretical Aspects of Reasoning About Data. Dordrecht: Kluwer Academic Publishing. ISBN 0-7923-1472-7, (1991)

[18] Revett, K., Analysis of a dobutamine stress echocardiography dataset using rough sets, Rough Sets and Intelligent Systems Paradigms RSEISP 2007 (LNAI 4585), Poland, June 2007, pp. 756–762.

(Kenneth Revett) University of Westminster, Harrow School of Computer Science,
London, England HA1 3TP
E-mail address: revettk@westminster.ac.uk

(Florin Gorunescu) Department of Mathematics, Biostatistics and Computer Science,
University of Medicine and Pharmacy of Craiova, Romania
E-mail address: fgorun@rdslink.ro

(Abdel-Badeeh Salem) Ain Shams University, Faculty of Computer Science and
Information Systems, Cairo, Egypt
E-mail address: abmsalem@yahoo.com

(El-Sayed El-Dahshan) Ain Shams University, Faculty of Computer Science and
Information Systems, Cairo, Egypt
E-mail address: e_eldahshan@yahoo.com