Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008

AN IMPROVED FEATURE EXTRACTION APPROACH BASED ON ROUGH SETS FOR THE MEDICAL DIAGNOSIS

WEI JIANG, YI-JUN LI, XIU-LI PANG

Information Management Research Center, Harbin Institute of Technology,
Harbin, 150001, P.R.China
E-MAIL: jwSeaBreeze@hit.edu.cn

Abstract:

This paper presents a novel approach based on Rough Sets to extract the complicated features from the medical diagnosis corpus. Some symptoms or basic features in the medical diagnosis are usually correlated. In general, the combinations of several basic symptoms may represent the disease more precision. However, the overmuch feature can reduce the generalization ability, or even many unfit features as the noise can decrease the model's performance. This paper proposes to apply the rough set theory to mine the complicated features, even from noise or inconsistent corpus. Secondly, these complex features are added into the Maximum Entropy model or Support Vector Machine etc. as a new kind of features, consequently, the feature weights can be assigned according to the performance of the whole model. The experiments in the Liver-disorders repository show that our method can improve the Maximum Entropy model by the precision 3.51%, improve the Support Vector Machine model by the precision 3.05%, improve the Naïve Bayes model by the precision 3.59%, and improve the Bayes and GoodTuring model by the precision 3.59%.

Keywords:

Medical Diagnosis; Rough Sets; Maximum Entropy Model; Support Vector Machine; Feature Extraction

1. Introduction

Medical Diagnosis (MD) is performed in order to extend the information beyond what is usually obtained from physical examination and medical history of a patient. The added information allows the physician to obtain a deeper insight of the patient's medical condition. Medical informatics has become an integral part of successful medical institutions [1].

The Medical Diagnosis system can provide two functions, including the primary diagnosis and the history cases. It can bring an aided reference for the doctors and the patients. The main methods can be divided into three categories. One is the rule based method, which make use of the medical knowledge and deduce the diagnosis result. However, not only it is exhausted to build the medical knowledge, but also it is difficult to represent the knowledge well and to perform the deducing process. The second category is statistic based method, which can automatically acquire the statistical knowledge, typically from the corpus. However, the sparse data problem is suffered from by the supervised learning model. The third category is hybrid method, which incorporate the statistic information with the medical knowledge. The medical knowledge is introduced in order to improve the generalization ability of model by means of fusing additional medical knowledge, usually, which is difficultly acquired from the training corpus.

Rough set theory has been developed in the field of medical informatics. Examples include data analysis of preoperative information of patients with duodenal ulcer treated with highly selective vagotomy [3][4], induction of prognostic rules by analyzing a database on women with breast cancer to determine short-term and long-term follow-up survival, development of a prototype expert system for assessing preterm birth risk [5], and diagnosis of progressive encephalopathy in children [6], and acute appendicitis [7]. The literature also documents methods that were developed to optimize the diagnostic process. Some examples of such methods are the analytical hierarchy process (AHP), a multi criteria decision making method that assists physicians in sequentially selecting the most appropriate tests for medical diagnosis [8]; use of modified Bayes' formula to analyze the sixteen test combinations for the diagnosis of Hepatolenticular Degeneration or Wilson's disease [9]. Evolutionary algorithms designed to improve the efficiency of healthcare institutions can be found in applications such as DIAPRO-diagnostic process [10].

In this paper, we focus on the complicated feature extraction. The feature distribution is usually complies with the Zipf's law, so the sparse data problem is hardly overcome, which is suffered from by the supervised learning model. In order to relax this difficulty, two kinds

978-1-4244-2096-4/08/$25.00 ©2008 IEEE

work are effective: 1) Introduce additional medical deducing knowledge; 2) extract and utilize effective features. The hybrid method attempts to improve the performance by introducing additional medical resource. On contrast, mining the complicated features is another effective approach to overcome the sparse data problem and improve the classifier's performance.

The complicated feature is helpful for the diagnosis decision. For instance when we know that the value of aspartate aminotransferase is 20 and the value of gamma-glutamyl transpeptidase is 14, we have more confidence to judge the liver disorder. However it is difficult to mine the complicated features, since there is the noise or the inconsistent sample in the corpus. This paper seeks to solve the problems mentioned above, by applying rough sets to automatically mine rough rules. And these mined rules can be used independently or integrated with other statistical techniques. Besides, considering each rough rule has the different effect in making decisions, we fuse the rough features into Maximum Entropy model (MEM) or the Support Vector Machine (SVM) classifier as a new kind of feature, accordingly, the proper weight of each rough feature is calculated according to the whole classifier's performance. Our method is verified in several well known evaluation corpora, and the experiments indicate its effectiveness.

The paper is organized as follows. Section 2 presents the complicated feature extraction based on rough sets. In Section 3, we present the utilization of rough rule features in the Maximum Entropy, Support Vector Machine and the Naïve Bayes classifier. In Section 4, we present the experiments and analysis, and Section 5 gives the conclusion and future research problems.

2. Complicated Feature Extraction

Feature space is an important factor to affect performance. Appropriate features can well represent decision knowledge, on the contrary, noise feature would be probably introduced and decrease the model performance.

In Medical diagnosis, it is important to extract complicated features to judge a patient's illness, for some symptoms or basic features in the medical diagnosis corpus are usually correlated. In general, the combinations of several basic symptoms may represent the disease more precision. For example, "cough" or "snivel" is symptom feature alone to judge if a patient has a cold. If they are combined as a feature, they play a better role. That is, if a person is coughing and sniveling at the same time, he has a higher probability to be judged as "cold".

In rough set theory, knowledge is represented via relational tables. An Information System can be defined as follows: $I = (U, A, V_a, f_a)_{a \in A}$, where U is a non-empty set of objects; A is a non-empty set of attribute a's; for each attribute $a \in A$, there is an attribute value V_a set and an information function $f_a : U \to V_a$. An equivalence θ on set U is called an indiscernible relation, and lower approximation for an object set $X \subseteq U$ is defined as $\underline{X}\theta = \{\theta x : \theta x \subseteq X\}$. However this formula is too strict to fit the requirement of NLP. The concept of α *-approximation* is provided: $\underline{X}\theta(\alpha) = \bigcup \left\{ \theta x : \frac{|\theta x \cap X|}{|\theta x|} \geq \alpha \right\}$, where α is an external parameter [11].

When extracting features, α *-Approximation* will probably cause the unbalanced support, since each selection of the decisions maybe has disproportion distribution. In order to let all the features added in can provide more evidences toward the right decision value, α *-Approximation* is introduced in our model [13]. Let filter parameter $\alpha_d \in [0,1]$, and the n-order rough rule set of keyword t is notated as R_t^n, then $R_t^n \in G_{t,n}$, and defined as: $R_t^n = \left\{ r \in G_{t,n} \mid r \in \underline{X}_{d,\theta}^{(i)}(\alpha_d) \right\}$, where $n = \left| A_f \right| - 1$, $i \in [1, K]$ and $G_{t,n}$ represents generalized LIT. In $G_{t,n}$, indiscernible objects are merged, the objects of each equivalence classes are counted and potential rule precision are calculated. If let each α_d is the same value, namely, let $\alpha_d = \alpha$ to all the decision attribute d, then λ *-Approximation* will back to the conventional definition of α *-Approximation*.

Table 1. 1-order MDIT liver-disorders

Condition attribute	Decision attribute
mcv	0
alkphos	1
sgpt alamine	0
sgot	1
gammagt	0
drinks	0

Medical diagnosis information table (MDIT) lists all the objects, which are potential features to assist in deciding how to do deal with noisy or redundant medical database. After merging indiscernible objects and counting each equivalence classes, α *-Approximation* is adopted to reducing data and extract rough rules.

Take 1-order MDIT for example, we suppose a Physician judge if a liver is ordered or disordered through the 10 symptoms and criteria, which has been mentioned above.

Firstly, let the tag of "disordered" be 0, and the tag of "ordered" be 1. Then collect the objects into MDIT (as Table 1 shown). After all the corpus samples are collected, we need to construct the generalized MDIT.

The rough rule count and precision needs to be calculated, which presents in the column Count and Precision when 2-order MDIT is made. Count in Table 2 is additive attribute of decision – making object, and it is used to statistic the count of the object which in the training data.

As the method of the feature extraction, it is need to be considered that if the algorithm can extract the distinguished and stable feature. Showed as Eq.(1) , x is objecr, $p(x \mid A_f(x))$ is the expected likelihood estimation restricted to x .

$$p(x \mid A_f(x)) = \frac{Cnt(x)}{\sum_{\{y:y\in GLIT, y\theta A_f(x)\}} Cnt(y)} \qquad (1)$$

The rough set rules can be extracted from Medical diagnosis information table (MDIT) according to α . According to Eq.(2), the extracted rules has the distinguished quality. Furthermore, we make the count more than the threshold based on the hypothesis : the rough set rules which do not appear often may not be stable. And it ensures the stability of the extracted rough set rules.

So, we ensure the stability and the quality of distinguished by extracting features though α . From other point of view, the precision of rules explains the contribution of the restriction relation (the degree of supporting the class attribute), then the phenomena of the uncertainty in the medical diagnosis can be dealt with.

Table 2. 2-order MDIT liver-disorders

Condition attribute	Decision attribute	Count	Precision
Mcv + alkphos	0	3	0.3
Mcv + sgpt alamine	0	20	0.9
Alkphos + sgpt alamine	1	5	1.0
Alkphos + gammagt	0	2	0.2
Alkphos + drinks	1	12	0.5
Sgot + gammagt	1	15	0.85
Gammagt + drinks	0	8	1

As shown in Table 2, α -Approximation can be used to acquire the rough rules. When set proper threshold parameters (such as Count > 10 and Precision = 0.5), "cv + sgpt alamine", "Alkphos + drinks" and "sgot + gammagt" as effective rough rules can be acquired. Note that position information is usually helpful, rules in left or right side is distinguished in our model.

3. Features Fusion and Expansion

As mentioned in section 2, physicians consider each medical diagnostic case as an individual one that involves the capturing and analyzing of different information classes regarding a patient's Health condition, obviously, the decision problem always is important to the diagnosis for the physicians. Decision problem in Medical Diagnosis is regarded as classifying problem, which is defined over $H \times T$ in Maximum Entropy model, where H is the set of possible medical diagnosis database around target name of illness that will be tagged, and T is the set of allowable tags. Then the model's conditional probability is defined as [13]:

$$p(t \mid h) = \frac{p(h,t)}{\sum_{t'\in T} p(h,t')} \qquad (2)$$

$$\text{where} \quad p(h,t) = \pi\mu \prod_{j=1}^{k} \alpha_j^{f_j(h,t)}$$

It has been pointed out that two kinds of decisions were dealt. One is the simple two categories problem, such as "disordered of a liver" and "ordered of a liver" are tagged as "0" and "1". The other is multi-categories problem. Two consecutive medical diagnosis decisions can be classified, and the tags are assigned from 0 to 3, such as "medium" is tagged as "3". H includes the near context and long distance context.

Rough rule features are defined as:

$$f_j(a,b) = \begin{cases} 1 & if((w = KeyWord) and (A_f(r) = b) \\ & \quad and (a = d)) \\ 0 & \end{cases} \qquad (3)$$

where the formula $A_f(r) = b$ represents that the conditional attribute of r can be reconstructed in the medical diagnosis database , and $a = d$ represents the decision attribute of d is equal to the tag of illness' names. For instance, ("sgot = 50" →"disordered") is rough rule feature, let the tag of "disordered" is 0, if there is "sgot = 51", then the Maximum Entropy model increases more evidences toward tagging 0.

In Medical Diagnosis (MD), the kernel task is to seek the mappings $\mathbf{x} \mapsto f(\mathbf{x}, \alpha)$, $\mathbf{x}$ be the input medical diagnosis features vector, y as the corresponding output tag of the illness (More has been detailed in our paper [13]).

In addition, high-order and lower-order rough rule

features worked together and are fused into Maximum Entropy model to overcome the data sparseness. This approach is based on the hypothesis that high-order feature is more accuracy but more sparser and that lower-order feature is less accuracy but less sparser.

4. Experiment and Analysis

The below two medical diagnosis data come from the UCI Machine Learning Repository: 1) Liver-disorder repository. It consists of 345 data. Among them, there are 288 training data, 57 test data. When applying rough set to extract complicated features, we select count 2 and rough set α parameter 0.8. There are 104 complicated features left. 2) Beast cancer repository. It consists of 286 data. Among them, there are 239 training data, 47 test data. When applying rough set to extract complicated features, we select count 2 and rough set α parameter 0.8. There are 28 complicated features left. When applying rough set to extract complicated features, we select count 1 and rough set α parameter 0.8, and obtain 5233 complicated features. The experimental results is shown in Table 3 and Table 4.

Table 3. Precision of open test in Liver-disorder repository

Diagnosis Model	Close test	Open test
Bayes	0.934028	0.456140
Bayes +RS	0.951389	0.526316
Bayes+GoodTuring	0.843750	0.561404
Bayes+GoodTuring +RS	0.861111	0.596491
ME	0.954861	0.526316
ME+RS	0.972222	0.561404
SVM	0.899300	0.561400
SVM + RS	0.930600	0.596500

Table 4. Precision of open test in Beast Cancer repository

Diagnosis Model	Close test	Open test
Bayes+GoodTuring	0.753138	0.680851
Bayes+GoodTuring+RS	0.736402	0.723404
Bayes	0.765690	0.680851
Bayes+ RS	0.736402	0.723404
ME	0.765690	0.702128
ME+RS	0.774059	0.702128
SVM	0.790800	0.702100
SVM + RS	0.790800	0.723400

The precision of the open tests in liver-disorder repository increase by Bayes, Bayes + GoodTuring, ME and SVM models. Bayes acquires more increase than the other two methods. The experiments in the Liver-disorders repository show that our method can improve the Maximum Entropy model by the precision 0.035088, improve the Support Vector Machine model by the precision 0.03051, improve the Naïve Bayes model by the precision 0.03587, and improve the Bayes and goodturing model by the precision 0.03587.

In Figure 1. and Figure 2., Bayes1 means the medical model is Naïve Bayes. Bayes2 means the medical model is Bayes + GoodTuring. As shown in Figure 1., before applying rough set α, Bayes + GoodTuring and SVM medical models have relatively high performances. After using rough set α to extract features, Bayes + GoodTuring and SVM have relatively high performance.

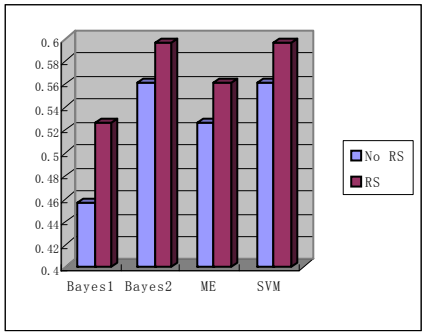

Figure 1. Comparison with the based RS(liver-disorder)

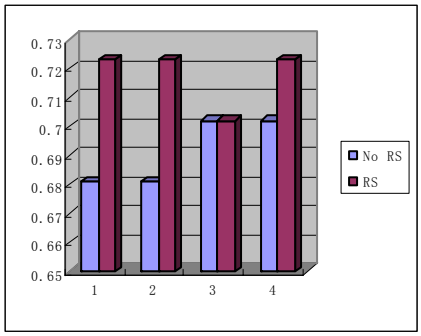

Figure 2. Comparison with the based RS(breast cancer)

As shown in Figure 2., in breast cancer repository, before applying rough set α, SVM has a relatively high performance. After using rough set α to extract features, Bayes and Bayes + GoodTuring have a relatively high performance. Bayes and Bayes + GoodTuring gets more increase than the other two methods.

Figure 3.[1] shows the relation between count and number of ME features. The number of MD features extracted by the rough set α in the four MD repositories

[1] "wdbc" represents the wdbc of breast-cancer-wisconsin repository in UCI repository, and "wpbc" represents then wpbc of breast-cancer-wisconsin repository.

exhibits the same trends. When count is growing, the number of MD features is decreasing. The figure shows that the number of features and the Count has a relationship of the inverse ratio.

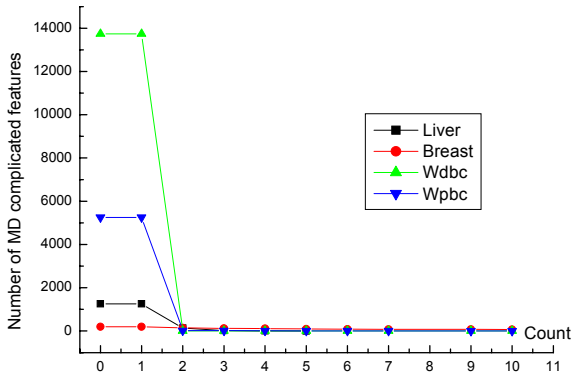

Figure 3. The relation between count and number of MD features

The following experiments explore the parameter influence to the precision of the classifier.

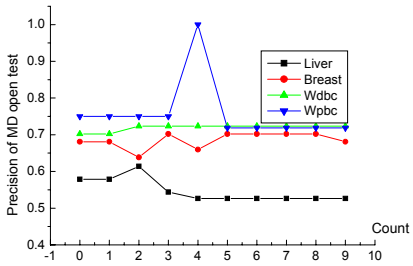

Figure 4. Pre. of open test when α is fixed

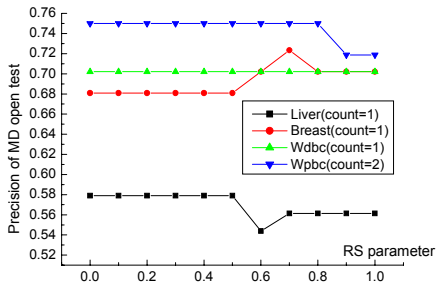

Figure 5. Pre. of open test when count is fixed

Figure 4. shows the precision of open test when α is fixed. In this case, the precision of open test fluctuates with the growing of count. When count gets a certain point, the precision gets the maximal point. Liver repository gets the best precision of open test when count = 2. Breast cancer repository gets the best precision of open test when count = 3. Wdbc repository gets the best precision of open test when count = 2-10. Wpbc repository gets the best precision of open test when count = 4. That is because when count is small, the instable and noisy features are imported. When count is too large, the extracted features will lose quality of stability for they are filtered too much useful MD features.

Figure 5. shows the precision of open test when count is fixed. When count is fixed, the precision of open test fluctuates with the growing of rough set α. When rough set α gets a certain point, the precision gets the maximal point. Liver repository gets the best precision of open test when α = 0.6. Breast cancer repository gets the best precision of open test when α = 0.8. Wdbc repository gets the best precision of open test when α = 0-1.0. Wpbc repository gets the best precision of open test when α = 0-0.8. When α is small, the instable and noisy features are imported, on contrast, a big α will bring a strict filter, so that a lot of qualified features will be cut off and abandoned. So the performance of the model and the scale of the model should be balanced. The selection of the proper parameters can extract stable and useful features, and the performance will be improved.

5. Conclusions

This paper presents a novel approach of Rough Sets to extract the complicated features in the medical diagnosis corpus, and these complex features are added into the Maximum Entropy model or Support Vector Machine etc. as a new kind of features, consequently, the feature weights can be assigned according to the performance of the whole model. And the experiments in the Liver-disorders repository show that our method can improve the Maximum Entropy model by the precision 3.51%, improve the Support Vector Machine model by the precision 3.05%, improve the Naïve Bayes model by the precision 3.59%, and improve the Bayes and goodturing model by the precision 3.59%.

The work in the future is concentrated on two respects. One is seeking more delicate feature templates according to MD knowledge to improve the precision of medical diagnosis, and dealing with them by the method proposed in this paper. The other is trying to incorporate class based feature, such as topic medical terms or those features that are extracted by the data-driven clustering algorithm.

References

[1] E. H. Shortliffe and L. M. Fagan, Medical Informatics: Computer Applications in Health Care and

Biomedicine, 2nd ed. New York, NY: Springer-Verlag. 2000

[2] H. E. Pople, "Heuristic methods for imposing structure on ill structured problems: The structuring of medical diagnosis," in Artificial Intelligence in Medicine. Boulder, CO:Westview, AAAS Selected Symp. 1982:119–185

[3] Z. Pawlak, "Rough sets: Theoretical aspects of reasoning about data," System Theory, Knowledge Engineering and Problem Solving, Kluwer, 1991, vol 9

[4] K. Slowinski,, "Rough classification of HSV patients," in Intelligent Decision Support: Handbook of Applications and Advances in Rough Sets Theory, System Theory, Knowledge Engineering and Problem Solving, vol. 11, Norwell, MA: Kluwer. 1992:77–93

[5] L. Woolery and J. Grzymala-Busse, "Machine learning for an expert system to predict preterm birth risk," J. Amer.Med. Informatics Assoc., 1994, 1(6):439–446

[6] A. Wakulicz-Deja and P. Paszek, "Diagnose progressive encephalopathy applying the rough set theory," Int. J. Medical Informatics, 1997, 46(2):119–127

[7] U. Carlin, J. Komorowski, and A. Øhrn, "Rough set analysis of patients with suspected acute appendicitis," in Proc. 7th Conf. Information Processing and Management of Uncertainty in Knowledge-Based Systems(IPMU'98). 1998:1528–1533

[8] F. Castro, P. Caccamo, J. Carter, B. Erickson, W. Johnson, E. Kesslar, N. Ritchey, and C. Ruiz, "Sequential test selection in the analysis of abdominal pain," Medical Decision Making, 1999, 19:178–183

[9] R. Richards, J. Hammitt, and J. Tsevat, "Finding the optimal multipletest strategy using a method to Logistic Regression," Medical Decision Making, 1996, 16:367–375

[10] V. Podgorelec and P. Kokal, "Towards more optimal medical diagnosing with evolutionary algorithms," J. Medical Systems, 2001, 25(3):195– 219

[11] Wang Xiaolong, Chen Qingcai, and Daniel S. Yeung, Senior Member, IEEE, Mining Pinyin-to-Character Conversion Rules From Large-Scale Corpus: A Rough Set Approach, IEEE transactions on systems, man, and cybernetics-part B: cybernetics. 2004, 34:834-844

[12] George-Peter K. Economou, Member, IEEE, Dimitris Lymberopoulos, Member, IEEE, Evy Karavatselou, and Costas Chassomeris, A New Concept Toward Computer-Aided Medical Diagnosis—A Prototype Implementation Addressing Pulmonary Diseases, IEEE Transactions on information technology in biomedicine. 2001, 5:55-66

[13] Wei Jiang, Xiao-Long Wang, Yi Guan et al. Applying Rough Sets in Word Segmentation Disambiguation Based on Maximum Entropy Model. Journal of Harbin Institute of Technology (New Series). 2006, 13(1): 94-98