Pattern Recognition Technique Based on Voting by Systems of Regularities

Senko O.V. [1] Kuznetsova A.V. [2],

1)Dorodnicyn Computer Center of Russian Academy of Sciences, Russia, 119991,
Moscow, Vavilova, 40,   senkoov@mail.ru

2) Institute of Biochemical Physics of Russian Academy of Sciences, Russia, 117997,
Moscow, Kosygina, 4,   azfor@narod.ru

New pattern recognition method is considered that is based on weighted voting by systems of subregions if feature space. Optimal subregion «syndromes» are searched with the help of optimal partitioning inside several families of different complexity levels. Results of experiments are represented.

Introduction. Pattern recognition methods based on collective solutions by ensembles of regularities is rather common approach now [1]. In this paper a technique is discussed that uses weighted voting by systems of «syndromes» - subregions in space of prognostic features where content of objects from one of the classes differs significantly from the contents of the same class objects in neighboring subregions. This technique is further development of method Statistically weighted syndromes (SWS) that was suggested in works [2,3].

Suppose that we try to recognize objects belonging to classes $K_1,\ldots,K_l$. At the initial stage for each of classes K_j set of «syndromes» $\tilde{Q}_j$ is calculated by some training information $\tilde{S}_o$. Suppose that object s^* is recognized by its vector description $\mathbf{x}^*$ that belongs to intersection of syndromes $q_1,\ldots,q_r$ from system $\tilde{Q}_j$. Estimate of object s^* for class K_j is calculated as prognosis of class K_j indicator function I_j at point $\mathbf{x}^*$. Prognosis is calculated with the help of statistically weighed voting procedure [4] :

$\Gamma_j(s') = \frac{\sum_{i=1}^{r} w_i v_i^j}{\sum_{i=1}^{r} w_i}$, where v_l^j - fraction of K_j among $\widetilde{S}_0$ with X from q_l, w_i is so called "weight" of the i-th "syndrome". At that "weights" maximizing special quasi-likelihood function [4] are used: $w_i = \frac{m_i}{m_i + 1} \frac{1}{\widehat{d}_i}$, where $\widehat{d}_i$ - estimate of variance of indicator function of class K_j indicator indicator function at "syndrome" q_i. Object s^* is placed to class $K_{j_{\max}}$ where $\Gamma_{j_{\max}}(s')$ is maximal value among $\Gamma_1(s'),\ldots,\Gamma_l(s')$.

"Syndromes" are constructed by optimal partitioning of prognostics features ranges. The partition R with maximal value of quality functional

$$F_g(R, \tilde{S}_0, K_j) = \frac{1}{v_0^j(1 - v_0^j)} \sum_{t=1}^{T} (v_l^j - v_0^j)^2 m_t$$

is searched for feature X, where T_R is number of subintervals in partition, m_t is number of objects from $\widetilde{S}_0$ with X from subinterval b_t, v_0^j - fraction of K_j among $\widetilde{S}_0$. SWS technique include selection of found regularities with the help of threshold for quality functional $F_g(R, \tilde{S}_o, K_j)$.

SWS technique demonstrate high effectiveness in many task but its evident drawback is incapability of effective recognition in case when syndromes can not be revealed or exactly described with the help with unidimensional partitioning only. Such syndromes may be discovered with the help of more complicated partitioning models. However as it was shown in [5] using of complex models may due to overfitting effect lead to "revealing" of false regularities that do not really exist in data. Including false regularities in ensembles actually may significantly decrease recognizing capability of collective solutions.

So it seems to be reasonable to use simultaneously several models with different levels of complexity. At that using of more complex model is justified when it allows significantly better separate descriptions of objects from different classes.

Multi-model statistically weighed syndromes. In present paper the pattern recognition method Multi-model statistically weighed syndromes (MSWS) is discussed that is based on the same voting procedure as SWS technique. “Syndromes” are searched with the help of four partitioning families: simplest uni-dimensional family with one boundary point (I); uni-dimensional with two boundary points (II); two-dimensional model with boundary lines that are parallel to coordinate axes(III); two-dimensional model with straight boundary lines that is arbitrary oriented with regard to coordinate axes (IV).

«Syndromes» in MSWS method are searched by optimal partitioning of corresponding features allowable region. In case of uni-dimensional models allowable intervals of single features are partitioned. In case of two-dimensional models optimal partitions allowable subareas of features pairs are searched. Instead of functional $F_g(R,\tilde{S}_o,K_j)$ used in SWS functional $F_l(R,\tilde{S}_o,K_j)=\frac{1}{\nu_0^j(1-\nu_0^j)}\max_{t\in T_R}[(\nu_t^j-\nu_0^j)^2 m_t]$

is maximized in MSWS method. The use of $F_l(R,\tilde{S}_o,K_j)$ instead of $F_g(R,\tilde{S}_o,K_j)$ is connected with necessity to measure qualities of partitions with different numbers of elements in the same units.

MSWS technique include selection of found regularities with the help of threshold Δ_l for quality functional $F_l(R,\tilde{S}_o,K_j)$. To diminish overfitting effect MSWS technique includes possibility to penalize regularities found with the help of more complicated partitions families. The quality functional $F_l(R,\tilde{S}_o,K_j)$ for families (II-IV) is multiplied by coefficient η. So threshold Δ_l and coefficient η are open parameters of method.

Experiments. Effectiveness of MSWS technique was evaluated in regard to SWS and some widespread pattern recognition techniques in four practical biomedical tasks.

1) Task of pneumonia severity computer diagnostics by 41 features in four groups with different severity levels evaluated by experts. The task was reduced to pattern recognition task with 4 classes. Full number of cases in initial dataset was 116.

2) Task of melanoma lesions diagnostics by features describing corresponding images. The task of diagnostics was reduced to pattern recognition task with 3 classes. Full number of cases in initial dataset was 80.
3) Task hysteromyoma relapse forecasting by set of immunological parameters. Forecasting of two severity levels by 69 immunological was reduced to pattern recognition task with two classes. Full number of cases in initial dataset was 60.
4) Task of forecasting of postoperative period severity. Forecasting of two severity levels by 11 preliminary selected features was reduced to pattern recognition task with two classes. Full number of cases in initial dataset was 331.

Three well known pattern recognition methods realized inside «RECOGNITION» systems [1] were considered besides SVM and MSWS: Fisher linear discriminant (LD), q-nearest neighbors (NN), support vector machines (SVM) with Gaussian kernel. In NN method optimal number of used nearest neighbors was chosen automatically in cross validation mode. The results of experiments are represent in Table 1. In upper section of each cell number of correctly recognized objects is given , in middle section – percent fraction of correctly recognized objects (P_s), in low section – average percent fraction by all classes (P_c .) The both values P_s and P_c are given because P_c more exactly describes situation than P_s when there is great difference between fraction of classes in training information. The results for SVM are given for best size of kernel from interval $[1.0, 5.0]$

Table 1. Results of experiments with several different pattern recognition techniques

	MSWS	SWS	FD	QNN	SVM
Task 1	78	76	64	49	69
	67.2%	65.5%	55.2%	42.2%	59.5%
	69.7%	65.9	49.3%	39.9%	54.5%
Task 2	54	51	48	44	51
	67.5%	63.8%	60%	55%	63.8%
	68.0%	63.7%	59.4%	55.3%	63.1%

Task 3	37	43	46	40	43
	61.6%	71.7%	76.7%	66.7%	71.7%
	70%	63.6%	51%	51%	51%
Task 4	212	170	205	220	216
	63.4%	51.4%	61.9%	66.5%	65.3%
	61.7%	52.2%	49.2%	61.1%	50.7%
Average P_c	67.35%	61.35	52.2%	51.82%	54.82%

Good performance of MSWS in terms P_c of may be notified . P_s values are higher for widespread techniques in tasks 3 and 4. But actually they classify objects to more numerous group.

Also we studied influence of coefficient η on MSWS performance at fixed value threshold Δ_l. Results are represented in Table 2. Values in cells exactly correspond to Table 1.

Table 2. . Results of experiments with MSWSW with several values of multiplier η.

$Task \backslash \eta$	0.1	0.3	0.5	0.7	0.9
Task 1	78	76	78	77	76
	67.2%)	65.5%	67.2%	66.2%	65.5%
	67.4%	67.0%	69.7%	64.2%	64.0%
Task 2	52	53	54	53	50
	65%	66.2%	67.5%	66.2%	62.5%
	65%	66%	68.0%	66.4%	63%
Task 3	33	41	37	39	39
	55%	68.3%	61.6%	65%	65%
	56.2%	71%	70%	75%	72%
Task 4	194	190	212	204	197
	58.6%	57.4%	63.4%	61.6%	59.5%
	57.6%	56.2%	61.7%	60.5%	58.0%
Average P_c	61.55%	65.05%	67.35%	66.5%	64.25%

It is seen that average P_c values are greater for η values 0.3. 0.5, 07 than for P_c values equal 0.1 or 0.9.

Conclusion. The experimental results demonstrated performance of MSWS method in regard to SWS and some widespread pattern recognition techniques. Higher effectiveness at multiplier values 0.3. 0.5, 07 says both about positive effect of more complicated regularities and necessity of more strict criteria for their selection.

Acknowledgements The work was supported by RFFI grants 08-01-90016, 08-07-00437.

References

1 Zhuravlev Yu. I. Ryazanov V.V., Senko O.V RECOGNITION. Mathematical methods. Program system. Applications (in Russian). -Moscow: Fuzis, 2006.

2 Kuznetsov V.V., Senko O.V., Kuznetsova A.V. et all. Recognition of fuzzy systems by method of statistically weighed syndromes and its use for immune and hematologic norm and chronical pathology.//Chemical Physics, 1996, v.15, №1

3 Jackson A.M., Ivshina A.V., Senko O.V., Kuznetsova A.V., et all. Prognosis Intraversical Bacillus Calmette-Guerin Therapy for Superficial Bladder Cancer by Immunological Urinary Measurements: Statistically Weighted Syndromes Analysis// Journal of Urology.-1998, -V. 159, -.№ 3, -P. 1054-1063

4 O.V. Sen’ko The Use of a Weighted Voting Procedure on a system of basic sets in prediction problems../Comp. Maths. Math. Phys. Vol. 35 , No. pp.1249-1257, 1995

5 Oleg V.Senko and Anna V. Kuznetsova, The Optimal Valid Partitioning Procedures . Statistics on the Internet ***http://statjournals.net/***, April, 2006