Research on Image Clustering Algorithm Based on Support Vector Machine and Hmm Model

Dong An

Xi'an International University, Xi'an, Shaanxi, 710077, China

Keywords: Support Vector Machine, Hidden Markov Model, Image Clustering Algorithm, Random Field

Abstract: With the progress of science and technology, the image data is increasing rapidly, including network image, video, etc.. How to manage a large number of image data effectively is a challenge for researchers. Image clustering algorithm is used to manage the image data and classify data information. In this paper, we propose an image clustering algorithm based on support vector machine and hidden markov model. Markov model can establish the mathematical relationship between image pixels in different layers, and we can predict the clustering result and the clustering center. Support vector machine is a new kind of machine learning algorithm based on statistical learning theory, which can find the optimal classification hyperplane in high dimensional feature space through optimum solution. Based on the basic theory of support vector machine, a remote sensing image classifier based on support vector machine is established. Experimental results show that the accuracy of image classification using svm is obviously superior than that neural network based algorithm and maximum likelihood algorithm.

1. Introduction

With the Rapid Development of Information Technology, a Large Amount of Multimedia Information, Especially the Image Information Has Emerged. the Large Amount of Data of Image Information, the Diversity of Content and the Characteristic of Non-Structure Have Seriously Affected the People' Effective Use of Image Information. So the Clustering of Massive Image Information Has Become an Important Research Direction of Image Information Management.

Traditional Image Processing Parallelization Method is Mainly Focused on Multi-Core Computing. Since 1990, Scientists Applied Artificial Neural Network to Image Classification, and Make Great Progress. Nevertheless, There Are No Reliable Rules for the Structure of the Neural Network. in Order to Improve the Accuracy of Image Classification, It is Necessary for Us to Find a New Algorithm for Image Clustering.

As a New and Effective Statistical Method, Svm is Widely Used in Pattern Recognition and Machine Learning, Because of Its Following Features:

- with Small Sample Learning.
- With anti noise performance.
- High learning efficiency.
- Wide applicability.

SVM algorithm is a convex secondary optimization problem, which can figure out the optimal solution, i.e. global optimal solution. In this paper, we establish a SVM model with high efficiency.

Hidden Markov Model is developed on the basis of Markov chain. Compared with Markov chain, HMM is more complicated, and observed events and states are not a one-to-one correspondence, but a set of probability distribution. HMM can describe the short time stationary and non-stationary features of signal, and handle sequence signals of different lengths. Therefore, HMM is widely used in time series analysis, speech recognition technology, and image clustering analysis.

  DOI: 10.25236/iiicec.2019.058

Because the different image data sets have different characteristics, and the purpose of the analysis is different, so there are many kinds of image clustering methods. In recent years, clustering technology is broadly divided into four categories:

- Clustering algorithm based on hierarchical.
- Clustering algorithm based on objective function.
- Clustering algorithm based on graph theory.
- Clustering algorithm based on density and network.

Image clustering and retrieval are two important research directions of the high level semantic understanding of computer vision, and the spatial information of image plays an important role in image feature modeling. In this paper, we apply SVM model and HMM to analyze image clustering method.

2. The Proposed Methodology

2.1 Support Vector Machine

In order to reduce the effect of imbalanced dataset on Support Vector Machine classification performance, a new under -sampling algorithm based on the twice support vector machine is proposed for imbalanced data classification. For samples of majority class, this algorithm deletes the samples far from the classification hyperplane. And for samples of minority class, this algorithm use over -sampling algorithm to add new samples.

Given a data sample: $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_i,y_i)\}$, $x_i \in R^n, y_i \in \{1,-1\}$

SVM mainly proposed to construct a classification hyper plane to cut apart two different samples, in order to figure out the maximum classification interval and keep the minimum error rate. There is a decision function:

$\min\varphi(\omega)=\frac{1}{2}<\omega, \omega>c\sum_{i=1}^{l}\varepsilon_i$

$s.t.\ y_i(<w,x_i>+b) \geqslant 1-\varepsilon_i, \varepsilon_i \geqslant 0, i=1,2,\cdots,l$

By using Lagrange operator can we get the pairing question:

$maxW(a)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}*\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_i K(x_i,y_i)$

$s.t.\ \sum_{i=1}^{l}\alpha_i y_i=0, 0 \leqslant \alpha_i \leqslant c, i=1,2,\cdots,l$

Where $K(x_i,y_i)$is the kernel function, $K(x_i,y_i)=<\Phi(x_i), \Phi(x_j)>$ use non linear mapping $\varphi: R^k| \rightarrow F$ to map training sample from input space to other feature space. Finally, we can obtain decision function:

$f(x)=sgn(\sum_{x\in SVs}\alpha_i y_i K(x_i,x)+b)$

2.2 Distance from Point to Hyperplane

The distance from sample x to classification hyperplane can be calculated as follows:

$$d(x)=\frac{\omega}{\|\omega\|}(x-x_0)=\frac{1}{\|\omega\|}(\omega x-\omega x_0)=\frac{1}{\|\omega\|}[(\omega x+b)-\omega x_0+b)]$$

Where x_0 is mapping of sample x, ω is normal vector of hyperplane, $\|\omega\|$ is the second order norm of ω. If $f(x_0)=\omega x_0+b=0$, then we have:

$$d(x)=\frac{\omega x+b}{\|\omega\|}$$

In terms of SVM derivation process, we can figure out that $\omega=\sum_{x\in SV} y_i a_i x_i$; Thus, for linear separable question, the distance from sample x to hyperplane is:

$$d(x)=\frac{\sum_{x\in SV} yi\alpha i K(xi,\ x)}{\|\omega\|}$$

Distance from class to classification hyperplane $D(c_i)$ is distance from class c_i to classification hyperplane, which is calculated as follows:

$$D(c_i)=\frac{1}{ni}\sum_{x\in C} d(x_j)$$

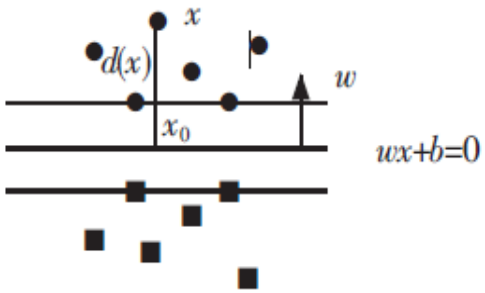

Fig.1 Distance from Sample Point to Hyperplane

2.3 Support Vector Machine Based on under-Sampling

There is a large amount of redundant information or helpless information for classification in most of the samples(such as samples far from classification hyperplane). These redundant information lead to the imbalance of the training samples, and then affect the final classification performance of the separator. So a common method is to remove these redundant information by using a certain strategy, that is under-sampling, such as DROP, CNN, clustering algorithm. However, these methods delete some boundary samples as well. In this paper, we propose under-sampling algorithm based on sample-classification hyperplane, the algorithm is described as follows:

- Aiming at training data set T, trained by using Support Vector Machine(SVM), and get classification hyperplane f(x), normal vector ω, support vector set *SV*, and corresponding coefficient of each *SV*.
- Compute the distance $d(x_j)$ between samples and classification hyperplane.
- Compute the distance $D(c_i)$ between class and classification hyperplane.
- For most classes of samples, in terms of given control parameters *a*, delete sample point $d(x)>a*D(c_i)$, and get new training set T'.
- Train T', if the classification effect is reached to the ideal state, then we can get final classification hyperplane and decision function; Otherwise, reset control parameters *a*, back to step 4
- Use interpolation method, increase samples.

Control parameters *a* is used to control delete the proportion of most samples, its value determined in terms of the ratio of minority samples amount and majority samples amount. That is $a=\frac{kni}{nj}$, where n_i is minority samples amount, n_j is majority samples amount, *k* is constant.

2.4 Hidden Markov Model

Aiming at image clustering algorithm, we should not only consider about image gray scale information, but also consider about the spatial constraint information, especially in the case of large amount of noise exited in an image. MRF algorithm with ability of anti-noise adequately considers neighborhood information. Therefore, we use MRF mode to describe spatial information, and we construct Markov space constraints.

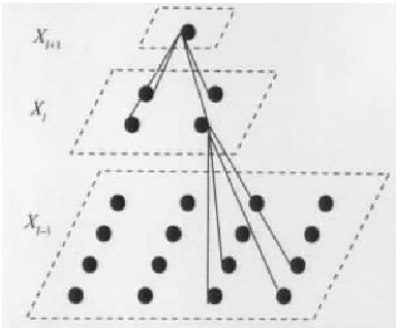

Fig.2 Sequence of Three Multi-Resolution Images Mapped Onto a Quadtree

Assume we have a Markov random field denoted as $Y=\{y_{ij}\,|\,(i,j)\in L\}$,where $L=\{(i,j)\,|\,i\in[1.M],\,j\in[1,N]\}$, y_{ij} is discrete random variables and $\forall y_{ij}\in Y, y_{ij}=k,\ k\in R, R=\{1,...,K\}$. $y=\{y_{ij}\,|\,y_{ij}\in R,(i,j)\in R\}$ is an MRF mode. On the basis of Hammersley-Clifford theorem, we can figure out that Gibbs distribution of *Y* is:

$$P(Y=y)=\frac{1}{Z}\exp(-U(y))$$

where $U(y)=\sum_{c\in\{C\}}V_c(y)$ is energy function, $V_c(y)$ is potential function, and $Z=\sum_{y}\exp(-U(y))$ is normalization factor. If sub-cluster coefficient is 1, then the potential function can be defined as:

$$V_c(y)=\begin{cases}\delta(y_{ij}-y_{nm}), c=\{(i,j),(m,n)\}\\ 0\end{cases}$$

where $\delta(t)\begin{cases}1, t=0\\ 0\end{cases}$.

Space constraint conditions of MRF can be described as follows:

$$P(y_{ij}=k\,|\,y_{mn}=l,(m,n)\in L,(m,n)\neq(i,j))=P(y_{ij}=k\,|\,y_{mn}=l,(m,n)\in\eta_{ij},(i,j)\in L)$$

where $k\in R, l\in R, \eta_{ij}$ is the neighborhood of pixel y_{ij}. According to Eq.(6.7.8), we have:

$$U(y_{ij})=\sum_{\{c|(i,j)\in C\}}V_c(y_{ij})=\delta(k-t_1)+\delta(k-t_2)+...+\delta(k-t_8)$$

$$\text{and } P(k\,|\,t_{ij})=\frac{P(k\,|\,t_{ij})}{P(t_{ij})}=\frac{\exp(-U(y_{ij}))}{\sum_{k\in R}\exp(-U(y_{ij}))}$$

thus, Markov space constraint is denoted as follows:

$$P=\{P(k\,|\,t_{ij})\,|\,(i.j)\in L, k\in R\}$$

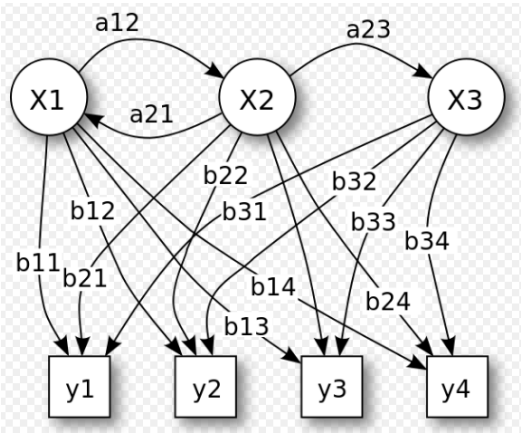

Fig.3 Hidden Markov Model

3. Conclusion

Image clustering algorithm is used to manage the image data and classify data information. However, the large amount of data of image information, the diversity of content and the characteristic of non-structure have seriously affected the people' effective use of image information.

Thus, how to manage a large number of image data effectively is a challenge for researchers. In this paper, we propose an image clustering algorithm based on support vector machine and hidden Markov model. SVM algorithm is a convex secondary optimization problem, which can figure out the optimal solution and HMM can describe the short time stationary and non-stationary features of signal, and handle sequence signals of different lengths. These two methods are widely used in the field of computer vision, such as image clustering, image segmentation, image fusion and so on. A remote sensing image classifier based on support vector machine is established in this paper, and experimental results show that image classification based on SVM and HMM has high accuracy.

References

[1] Villalba, Luis Javier García, Ana Lucila Sandoval Orozco, and Jocelin Rosales Corripio. "Smartphone image clustering." Expert Systems with Applications 42.4 (2015): 1927-1940.

[2] Wang, Haoxiang, and Jingbin Wang. "An effective image representation method using kernel classification." 2014 IEEE 26th International Conference on Tools with Artificial Intelligence. IEEE, 2014.

[3] Wang, Jingyan, et al. "Image tag completion by local learning." International Symposium on Neural Networks. Springer International Publishing, 2015.

[4] Zhao, Cuimei, et al. "Pseudocapacitive properties of cobalt hydroxide electrodeposited on Ni-foam-supported carbon nanomaterial." Materials Research Bulletin 48.9 (2013): 3189-3195.

[5] Dai, D., & Van Gool, L. (2016). Unsupervised High-level Feature Learning by Ensemble Projection for Semi-supervised Image Classification and Image Clustering$. Technical report.

[6] Bi, Chujian, et al. "SAR image restoration and change detection based on game theory." Intelligent Computing and Internet of Things (ICIT), 2014 International Conference on. IEEE, 2015.

[7] Schneider, Caroline A., Wayne S. Rasband, and Kevin W. Eliceiri. "NIH Image to ImageJ: 25 years of image analysis." Nat methods 9.7 (2012): 671-675.

[8] Shi, Yuqing, Shiqiang Du, and Weilan Wang. "Local consistent low rank representation for image clustering." Control and Decision Conference (CCDC), 2016 Chinese. IEEE, 2016.

[9] Schilder, Paul. The image and appearance of the human body. Vol. 163. Routledge, 2013.

[10] Schindelin, Johannes, et al. "Fiji: an open-source platform for biological-image analysis." Nature methods 9.7 (2012): 676-682.