Cohesive Feature Selection for Classification and Retrieval of Data

Saurabh Mukherjee[1], Shashikala Tapaswi[2], Renu Jain[3]

Abstract

Feature selection plays a vital role in any problem related to pattern recognition. Various tools and methods are used for feature definition, feature analysis, feature classification, and feature cohesiveness. Much of the work has already been done in all these fields earlier baring a few. Feature selection and classification has undergone many challenges from last one decade but still no robust method has been discovered, which can create a nexus between the human perception and the machine intelligence. The basic and fundamental problem that lies with the semantic gap between two agents namely, human and machine is 'perception' of similar kind of information. In the present paper, an attempt has been made to understand the concept of perception and how to replicate the same in the machine in terms of 'Cohesive feature'. Selection and classification of feature has been undertaken with special reference to conditional and class conditional representations and independence. Experimental Results performed on features have shown promising results. The comparison of the proposed approach has been done with various existing independent features and the results are quite comparable.

Keywords : Cohesive Feature, Perception, class conditional, classification, Selection.

[1]Dept. of Computer Applications, PIM, Gwalior.

[2] Associate Professor, ABVIIITM, Gwalior.

[3]Professor in School of Mathematics & allied Science, Jiwaji University, Gwalior.

I. Introduction

Information of any types such as Archival storage (filing), finding (retrieving) of information (data) exists for several thousand years, and both go hand in hand as a fundamental part in human nature(Peter Kovesi,2005). Human quest for information retrieval has came across many years. The amount of information grew continuously. Thus, in the earlier years, looking for some specific information implied searching manually through data even though indexing technique already existed(Peter Kovesi,1999). The skilled used of information searching was still a major concern. Exhaustive search was one way out right there. This means laborious and time-consuming work already for small data collection, not to mention larger ones. Besides searching by oneself, one could also look for or ask someone who knows this particular field to assist and find the desired information faster.

Somehow, this problem of manual work has been minimized in the today's Era with the advent of technology? In the new era, computers and electronic storage media helped to record data and information more efficiently a effectively such as in database systems.

Huge amount of information is being stored in the databases, which prompted for retrieving the information at a very fast rate.

Furthermore, this allowed the use of computers in which replacing the manual approach by an automatic search is being done. However, the automatic search is in need of a suitable retrieval interface(Enser,1995). User

expectation is always more than what system can provide. Although the manual search is slow, but it has the advantage, a human is conducting it and vice versa for the automatic search, i.e., it is fast but lacks the main concept of semantic knowledge, which only humans can provide and understand. Two challenges come with the user interaction of the electronic retrieval system compared to searching conventional databases. The first may be described as fuzziness because the user does not precisely know how to express the information he/she is looking for. Hence, queries include vague conditions(Eyre,1998). The second may be circumscribed as uncertainty where thesystem does not have the knowledge about interpreting the content of the data. Thus, this leads to inaccurate and missing results. Therefore, the retrieval system has to provide a user friendly interface to support the user in its needs for an efficient retrieval (Ballard et.al, 1991).

The paper is organized as follows Section 2 discusses the role of feature selection and classification, Concept of conditional independence is mentioned in Section 3. An element of class conditional representation with special reference to cohesive feature is described in Section 4. The details of proposed extended algorithm of Marco have been done with Cohesiveness features are given in Section 5 followed by experimental results and further discussion in Section 6.

2. Role of Feature Selection And Classification

Feature selection and classification here plays a pivotal role in this regard. Features, which are inherent in any object under study, is the integral part of any retrieval process, be it a text or an image or audio or video data. Feature extraction thus plays a major role in the searching process. Feature classification too deals with various independent as well as dependent concepts, which can be applied to get the desired information in the retrieval systems.

World Wide Web as the name specifies is a collection of trillions and trillions of the digital information. With increasing computerization, more and more data are available and stored digital form, which made it necessary to search huge databases and data collections such as the WWW more Cohesive Feature Selection for Classification and Retrieval of Data Saurabh Mukherjee, Shashikala Tapaswi *Member IEEE*, Renu Jain I2 efficiently and effectively. At the beginning of the WWW, there were no rules present for the definition of content and how to handle it. Over the years, the WWW changed to information medium and searching those information, especially text, became easier due to search engines such as Yahoo, MSN, and Google(Graham et.al, 1998).

However, information retrieval by text is not as easy as one might think because it is more than just matching words, phrases, or sentences. The issue here is that words can have different meanings in different correlations such as homograph (word written in the same way with different meaning) and synonym (words have the same meaning but written differently). How often does it happen that one searches for something specific in the WWW by using one of those search engines but does not get any relevant results, only the words seem to match but not their actual meaning of context one had in mind (Stix,1997). This observation is of its own credit; because to understand the concept of features, it is most important to understand its locality of references. The retrieval process merely use the concept of data retrieval rather than information retrieval as required. Now, that data is more vague than information is easy to understand. As per our own need, data is transformed into information.

Therefore, when TBIR is used it is not the information (with semantics) that is matched, it is the data (symbolic representation of words), that is matched, thus arising a lots of problem in the query domain due to its fuzziness and probabilistic nature (Frankal et.al,1996). Due to the aforementioned fuzziness and uncertainty , text based retrieval is not semantic which means that a search engine, a system, or a machine does not know what these symbols, character or numbers mean. Therefore, it is mainly just a data (symbol) matching and hardly a semantic content matching by those search engines even though retrieval and its results have improved over the past years but are still far from optimal (due to the lack to understand human perception).

(Stephan Ullman) in his research has rightly referenced that besides text, huge interest grows in searching multimedia content such as images and videos due to increasing digitalization based on the soaring number of digital devices (mobile phone, camera, etc.), which capture and store personal multimedia content. Moreover, the analogue audiovisual content from earlier ages is also converted into digital form. Hence, there is a demand to effectively store and organize such digital collections to support efficient queries and access schemes during retrieval, no matter if these collections are private or public.

However, this type of search has the same challenge as the text retrieval, i.e. lack of semantics or more precisely how to understand and describe the semantic content of a digital item? In order to accomplish this, two approaches can be considered: manual and automatic. For the manual case, the content description is performed entirely by humans.

Here, the content is described by so-called *Tags*, which are set by user(s) who provide(s) the item (e.g. images for Flickr, video for YouTube). Its advantage is that the image or video is provided with a semantic content description due the human interaction. But this also bears a couple of risks.

Firstly, to find an image or video, it has to have *Tags* and, secondly, these *Tags* have to provide a relevant and meaningful description. On the other side, the automatic case normally tries to describe the image content without any human interaction or intervention. Note further that these two approaches are not strictly separated and might be used simultaneously, especially if accurate and pre-processing steps such as segmentation may provide meaningful regions (objects). Moreover, besides describing the actual regions better based on their local properties (features), this further allows the integration of region relationships into the content description (Enser et.al, 1995). However, note that this may provide a better content description but will still lack the semantic meaning of the regions.

3. Concept of Conditional Independence

Let us take X and Y be any random variables having arbitrary value say (λ). Let the joint density of (X,Y) be p(x,y) , (X,Y)'s marginal density be p(x) and p(y) and the conditional density of (X,Y) be p(x/y).Here, if (X,Y) are independent than p(x,y)=p(x)p(y) and p(x/y)=p(x). The definition of independence can be extended to a multivariate case (X1, X2 ...XN) as p(x) = p(x1)...p (xN).

(Bressan et.al,2003) in his paper represents a global independence of conditional independence.In this regard simpson's paradox is know as the most versatile counter example in the regard.The falseness of the implication can also be visualized considering the random variables (X,Y) with retangular distribution in the square of

(lamda) =[0,1]x[0,1]. The random variable as defined by Bressan for the set {(x,y) ∈ (λ),x>y)}and 0 otherwise.

Here the role of conditional independence deals a great bit of information regarding any of the value of a selected feature. Given any random value say Z the information regarding Y can be obtained with the help of existing knowledge of the instance of Y(Ballard et.al,1991). In the other words, we can say that given Z, X -> Y or Y can be functionally determined by X for any random variables.

4. Elements of Class Conditional Representation With Special Reference to Cohesive Feature

The case in which class-conditional independence is encountered has interesting consequences in the field of statistically pattern recognition. Given a class of K=(λ) {C1...Ck), a set of features represented by an N-dimensional random vector say x= (x1,....xn), the Maximum a Posterior (MAP) and the Maximum Likelihood (ML) solutions both use the concept of class – conditional densities $p(x/C_k)$. If the densities are independent and equiprobabilities are taken into account then according to the Naïve Bayes rule the following equality given by Vitria[4][14] holds good.

C_{naive}= arg max(k=1....k) $\prod$ (n=1....N) $P(x_n/c_k)$ (1)

The problem of feature selection for the classification can be stated as, given a set of features representing our data, select a subset such that reduced sets work better than the comprehensive sets of features. Class seperability is an important criterion to choose a set of features (Bearman et.al, 1999). To categories each subsets of features from the entire database of features require loads of computation, resulting it to convert in a combinatorial problem.

Various feature extraction modes are used like that of Mahalanobis and Bhattacharya distances, guaasian divergence, fisher ratio test analysis etc. as given by K.Fukunaga. A commonly used distance measure is given by Kullback- Leiber distance in class condition representations KL (Ci, Cj)= $\int$ (λ) p(x/Ci) log [p(x/Ci)/p(x/Cj)] dx (2)

where 1<=I, j<=K.

Dcohesive(Ci, Cj)= $\int\int$ (λ) p(x/Ci) p(y/Ci) log [p(x/Ci)/p(x/Cj)] dy dx

where 1<=I, j<=K. (3)

5. Proposed Extended Algorithm With Special Reference To Cohesive Feature

Algorithm:-

1. Treat each class as a set of independent class.
2. Let the classes be from k= 1....k.
3. Use computational complexity by measuring the mutually independent components Like x_k, Wk and v_k, where W_k and X_k are the projection matrix and v_k is the normalizing constant using (1)
4. If marginal densities are unavailable, project class samples $s=p_k(w_k m(x - \overline{xk}))$.
5. Compute the cohesive features using summation of the independent components.
6. $D_{cohesive}$= $\sum$ (i=1....n) (k=1....k) KL (Ci, Cj) using (2)
7. $D_{cohesive}$ (Ci,Cj)= $\sum$ ((λ) p(x/Ci) p(y/Ci) log [p(x/Ci)/p(x/Cj)])
8. Estimate $D_{cohisiveness}$ and Estimate pk (s_m)
9. End loop.

6. Experimental Results And Further Discussions

In a recent survey of feature selection, Jain and Zonkar performed an experiment using artificial two-class sample. The two classes have multivariate normal 50-dimensional distributions with covariance given by the identity matrix. The first d-features are known from the first instance .They proposes a measure of average quality for the feature selection criterion and we have varying the number of training classes using the cohesive feature measures.

The present experiment uses the above concept but using cohesive features, an intensive processing has been done firstly to select the features, and then to show their performance in terms of cohesiveness.

Features propose a measure of average quality for the feature selection criterion varying the number of training patterns per class(Besser et.al,1997). The maximum possible value for this average quality is one, meaning that the 50 possible feature subsets were the optimal subset for the ten data sets. Experimentally the divergence is a fairy robust criterion with performance.

The database consists of 100 of images (.bmp or .jpg). SQL Server database is used to store the images. VB.net is used as the front-end tool. All the images stored in the database which consists of Uniform resolution.

The class-conditional competency is done with the basis of cohesive feature named Naïve-bayes techniques. The existing problem tried naïve Bayes on 100,000 random 8- feature combinations for each class, giving classification of 83.17%.

The present experiment using cohesive feature uses class conditional feature with 16-feature combinations for each independent class, giving classification of 88.954%.This shows a promising result, which can be improved more in further research.

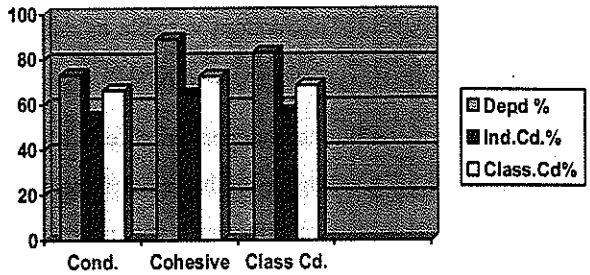

Figure 1 : Diagram Shows Three Factors for Feature Selection and Classification

The above chart shows the experimental results of using cohesive features for classification and selection. In fig (1) above, three parameters have been taken. These are shown in the above fig (1). The abscissa shows three parameters viz, Condition, Cohesive feature, and Class Condition. The ordinate shows the percentage of corresponding parameters retrieved. The comparative result shows the promising results as predicted. The class condition component can be taken independently as well as with intersection of the feature classes, which can be extended and will be the next step of this research.

References

[1] Peter Kovesi, *"Shapelets Correlated with Surface Normals Produce Surfaces"*, 10th IEEE International Conference on Computer Vision, Beijing, PP. 994-1001, 2005.

[2] Peter Kovesi, *"Phase Preserving Denoising of Images"*, The Australian Pattern Recognition Society Conference : *DICTA'99*, Perth WA, PP 212-217, Dec. 1999.

[3] Johansen. S. M. B, Laugesen. J. L, Dahl. C. K, Esbensen K. H and Frost, M. B.*AMT, "a second generation look: optimization of algorithm performance and systematics of method-modification"*, in preparation, 2004.

[4] Marco Brenssa and Jordi Vitria, *"On the selection and classification of independent features"*, IEEE TPAMI, Vol. 25, No.10, October 2003.

[5] Stephan Ullman, M.Sc thesis, *"Region based multimedia indexing and retrieval framework"*, Tampere University of Technology.

[6] Enser. P.G.B, *"Progress in documentation: pictorial information retrieval"*, Journal of Documentation, 51, PP. 126-170, 1995.

[7] Eakins. J.P, *"Techniques for image retrieval"*, Library & Information Briefings, 85. London: South Bank University, Library Information Technology Centre, 1998.

[8] Graham. M and Eakins. J, *"ARTISAN: a prototype retrieval system for trade mark images", VINE*, No. 107, PP. 73-80, 1998.

[9] Swain. M.J and Ballard D.H, *"Color indexing"*, International Journal of Computer Vision, 7(1), PP. 11-32, 1991.

[10] Frankel. C, Swain. M.J and Athitsos. V, *"WebSeer: an image search engine for the World Wide Web"*, Technical Report 96-14. Chicago, Ill.: University of Chicago, Computer Science Department, 1 August 1996, AltaVista Picture.

[11] Eyre. J, *"Distributed image services", VINE*, 107, PP. 65-72, 1998.

[12] Stix. G, *"Finding Pictures on the Web"*, Scientific American, 276 (3), March 1997, Blackaby. J and Sandore. B, *"Building integrated museum information retrieval systems: practical approaches to data organisation and access"*, Archives and Museum Informatics, 11, PP. 117-146, 1997.

[13] Bearman. D and Trant. J, *"Unifying our cultural memory: could electronic environments bridge the historical accidents that fragment cultural collections"*, In : Dempsey. L, Criddle. S and Hestletine. R, (eds.), *"Information landscapes for a learning society"*, London: Library Association, 1999 (forthcoming).

[14] Besser. H, *"Image databases: the first decade, the present, and the future"*, In : Heidorn. P.B and Sandore. B, (eds.), *"Digital image access and retrieval"*, Urbana-Champaign, Ill.: University of Illinois at Urbana- Champaign, Graduate School of Library and Information Science, PP. 11-28, 1997.

Author's Biography

Prof. Saurabh Mukherjee is actively engaged in research and development for more than 10 years. He is recipient of many awards for best research paper presentation in JSRS. His work on VIRTUAL REALITY has been published by IEEE Computer Society,USA.He is a reviewer of various international journals and conferences like IEEE Transaction on Fuzzy Systems,ITNG (USA), NISS (China) to name a few. He had been invited to give oral collage presentation in IIM, Ahmedabad. His active research area is in digital image processing, soft computing, advance operating system; advanced computer graphics etc. He has received one international scholarship from Chinese govt.

Dr Shashikala Tapaswi is an associate professor in ABVIIITM, Gwalior. Earlier she was associated with MITS,Gwalior.She has 21 years of teaching experience.Her research areas are Image Processing , Computer networks , Mobile and Ad-hoc Networks.

Dr. (Mrs.) Renu Jain is a Professor in School of Mathematics and allied Science, Jiwaji University, Gwalior. She is having more than 26 years of teaching and 27 years of research experience, respectively. She had been the guide of 08 Ph.D students. She was a recipient of Commonwealth Scholorship and done research in Imperial College,London.