บทที่ 2

หลักการ และทฤษฎีที่เกี่ยวข้อง

บทนี้กล่าวถึงหลักการและทฤษฎีของการจำแนกเอกสาร การสร้างดัชนีเอกสาร การลดมิติเอกสารโดยการการลดมิติโดยการเลือกคำสำคัญ การลดมิติโดยการสกัดคำสำคัญ ทฤษฎีโครงข่ายประสาท โดยมีเนื้อหาดังนี้

2.1 การจำแนกเอกสาร (Text Categorization)

เป็นกระบวนการจำแนกเอกสารอัตโนมัติโดยใช้เทคนิคของเครื่องมือสำหรับเรียนรู้ (Machine learning) กระทำโดยการจัดเอกสารลงกลุ่มที่ได้กำหนดไว้แล้ว เช่น เซต $C = \{c_1,..,c_m\}$ แต่ละกลุ่ม c_i โดยที่ i มีค่าเท่ากับ 1 ถึง m ซึ่งกลุ่มเหล่านี้จะถูกใช้ในการพิจารณาเอกสาร d_j ว่าควรอยู่ c_i หรือไม่ โดยตัวจำแนก (Classifier) เอกสาร $C_0 = \{d_1' ,...,d_s'\}$

กระบวนการนี้ต้องการข้อมูลข้างต้น $C_0 = \{d_1' ,...,d_s'\}$ ซึ่งเอกสาร d_j' โดยที่ j มีค่าระหว่าง 1 ถึง s ซึ่ง d_j' ต้องเป็นสมาชิกใน c_i ทำการแบ่งข้อมูล $Tr = \{d_1' ,.., d_g'\} \in C_0$ ซึ่ง Tr คือข้อมูลที่จะถูกนำไปใช้ในการเรียนรู้ โดยแบบจำลองสำหรับการจำแนกลักษณะเฉพาะของแต่ละ c_i ในการวัดประสิทธิภาพ สามารถทดสอบโดยการนำข้อมูลที่เหลือการแบ่งข้อมูลที่ใช้ในการเรียนรู้ของแบบจำลองนำมาทดสอบการจำแนก โดยกลุ่มสำหรับทดสอบ Ts ได้จากการนำข้อมูลที่เหลือจากใช้สำหรับเรียนรู้ $Ts = \{d_g' +1,..., d_s'\} = C_0 - Tr$ จากนั้นทำการทดสอบความถูกต้องของตัวจำแนกจากการวัดประสิทธิภาพการจำแนกของแบบจำลองบนข้อมูลสำหรับทดสอบ

(Caropreso, Matwin and Sebastiani, 2001)

2.2 การสร้างดัชนีเอกสาร (Document Indexing)

การทำดัชนีของเอกสารมีความจำเป็น เนื่องจากข้อมูลในเอกสารมีรูปแบบเป็นข้อความ (Text) ไม่สามารถนำมาใช้คำนวณได้ตรงกับตัวจำแนกหรือ ขั้นตอนวิธีสำหรับจำแนก (Classifier building algorithm) ดังนั้นจึงต้องมีกระบวนการแปลงเอกสาร d_j ให้อยู่ในรูปของตัวแทนเอกสารที่มีรูปแบบมาตรฐานที่สามารถนำไปใช้ในการเรียนรู้ และทดสอบได้ วิธีการสร้างดัชนีเอกสาร คือ การสร้างตัวแทนเอกสาร d_j ให้อยู่ในรูปแบบของเวคเตอร์น้ำหนักของคำสำคัญ (Vector of term weights) $d_j = (w_{1j} ,.., w_{|T|j})$ โดยที่ T คือ เซตของคำสำคัญ

doc / term	d_1	d_2	d_k
term1	w_1	...	w_{1k}
.	.	.	.
.	.	.	.
.	.	.	.
termn	w_{n1}		w_{nk}

รูป 2.1 แสดงการสร้างดัชนีเอกสาร

W_{kj} มีค่าเท่ากับ จำนวนของคำสำคัญ $term_k$ ในเอกสาร d_j การแทนเอกสารนี้คือการแทนเอกสารในรูปแบบของความถี่ของคำสำคัญ และวิธีหาค่าน้ำหนักที่เป็นที่นิยมวิธีการหนึ่งคือ ทีเอฟไอดีเอฟแบบมาตรฐาน (Standard tfidf) มีสมการดังนี้

$$tfidf(t_k, d_j) = \#(t_k, d_j).\, log\, \frac{|T_r|}{\#T_r(t_k)}$$

โดยที่ # (t_k, d_j) คือจำนวนครั้งที่พบ t_k ในเอกสาร d_j และ # $T_r(t_k)$ คือจำนวนเอกสารที่บรรจุ t_k อยู่ จากนั้นมีการนำเสนอการนอร์มาไลซ์ (Normalized) โดยโคไซน์นอร์มาไลเซชัน (Cosine normalization) โดย

$$W_{kj} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|} \left(tfidf(t_s, d_j)\right)^2}}$$

(Sebastiani, 2002)

2.3 การลดมิติเอกสาร (Dimensionality reduction)

ในการจำแนกเอกสาร เมื่อเราแปลงเอกสารให้อยู่ในรูปของดัชนีเอกสาร โดยคำสำคัญได้มาจากเอกสารสำหรับเรียนรู้ (Training data) พบว่าคำสำคัญที่ได้และถูกนำมาใช้ในการสร้างดัชนี เอกสารมีเป็นจำนวนมาก ซึ่งถ้ามองเอกสารในรูปของเวคเตอร์ พบว่ามิติของเอกสารหนึ่งๆ มีจำนวนมากตามไปด้วย อาจจะก่อให้เกิดปัญหาในการจำแนก เช่นคำสำคัญบางคำนั้นไม่ได้แสดงความแตกต่างของกลุ่มอย่างแท้จริง เมื่อนำคำสำคัญเหล่านั้นมาใช้ในการพิจารณากลุ่มของข้อมูลอาจทำให้ความถูกต้องในการพิจารณากลุ่มของตัวจำแนกลดลง เพื่อแก้ปัญหานี้ จึงได้มีการใช้กระบวนการต่างๆ ในการลดมิติของเอกสาร แต่ในการลดมิติของเอกสารในแต่ละงานต้องมีการหาจำนวนมิติที่เหมาะสมเพื่อลดปัญหาสองประการ คือ การเลือกจำนวนมิติที่มากเกินไป ซึ่งจะส่งผลให้ประสิทธิภาพต่ำลง ทั้งด้านความถูกต้องที่เกิดจากมีสิ่งรบกวนในการคำนวณ (Noise) และการใช้

เวลาในการจำแนก และการเลือกจำนวนมิติที่น้อยเกินไป อาจทำให้ตัวจำแนกไม่สามารถหาความแตกต่างระหว่างกลุ่มได้ดีเท่าที่มันควรจะเป็น การลดมิติเอกสารมีสองประเภท ประกอบด้วย การเลือกคำสำคัญบางคำมาใช้ในการสร้างดัชนีเอกสาร (Sebastiani, 2002), (Deerwester, Dumais, Landauer, Furnas, Harshman, 1990) ตัวอย่างการศึกษาและเปรียบเทียบกระบวนการการเลือกคำสำคัญ เช่นงานของ Yang (Yang and Pedersen, 1997) และงานของ Zheng (Zheng, Wu and Srihari) การเลือกคำสำคัญในการสร้างดัชนีเอกสารนั้นสามารถเพิ่มความถูกต้องของแบบจำลองการจำแนกได้ วิธีการเลือกคำสำคัญไม่ได้ใช้ในงานด้านการจำแนกเท่านั้น ยังมีการใช้งานในด้านการจัดกลุ่มด้วยดังเช่นงานของ Liu (Liu, Liu, Chen and Ma, 2004) การลดมิติเอกสารอีกวิธีหนึ่งคือการสกัดคำสำคัญเป็นวิธีการแปลงเวคเตอร์เอกสารให้อยู่ในรูปแบบอื่นที่มีมิติเล็กลง โดยวิธีการที่มีประสิทธิภาพและเป็นที่นิยมคือการสร้างดัชนีความหมายภายใน (Barbara, 2000) ตัวอย่างการใช้งานเช่นงานของ Cheng (Cheng and Soon, 2007) ได้มีการใช้กระบวนการแอลเอสไอ ในการสร้างตัวแทนเอกสารเพื่อเป็นข้อมูลเข้า ให้แก่โครงข่ายประสาท งานของ Wu (Wu and Gunopulos, 2002) ใช้เทอร์มเฟรซ (Term phrase) และดัชนีความหมายภายในในการเพิ่มประสิทธิภาพการจำแนกเอกสาร งานของ Zelikovitz (Zelikovitz and Hirsh,2001) ทำการสร้างสร้างดัชนีความหมายภายในบนดัชนีเอกสารที่ประกอบด้วยคำสำคัญและแบกกราวเท็กซ์ (Background text) จากงานวิจัยต่างๆ พบว่าดัชนีความหมายภายในสามารถเพิ่มความถูกต้องของการจำแนกเอกสารได้ การลดมิติเอกสารโดยการสกัดคำสำคัญอีกวิธีการหนึ่งคือวิธีพีแอลเอส เป็นกระบวนการหาพื้นที่ความหมายคล้ายกับกระบวนการดัชนีความหมายภายใน สิ่งที่แตกต่างจากดัชนีความหมายภายในคือ มีการใช้ข้อมูลของกลุ่มเข้ามาร่วมในการสร้างพื้นที่ความหมาย (Wang, Nie, 2003) โดยการลดมิติเอกสารโดยวิธีพีแอลเอสถูกนำเสนอโดย Zeng (Zeng, Wang and Nie, 2007) ผลที่ได้พบว่าพีแอลเอสช่วยให้แบบจำลองการจำแนกเอกสารมีความถูกต้องของการจำแนกสูงขึ้นและลดเวลาการคำนวณ

2.3.1 การลดมิติโดยการเลือกคำสำคัญ (Dimensionality reduction by term selection)

การลดมิติด้วยวิธีการเลือกคำสำคัญบางคำจากคำสำคัญทั้งหมด หรือเรียกอีกอย่างว่าการเลือกลักษณะเด่น สมมุติให้จำนวนคำสำคัญทั้งหมดคือ $|\tau|$เมื่อผ่านกระบวนการเลือกคำสำคัญแล้วจำนวนตัวแทนของคำสำคัญทั้งหมด $|\tau'|$ค่าของ $|\tau'|$จะมีค่าน้อยกว่า $|\tau|$มาก ($|\tau'| \ll |\tau|$)วิธีการที่ใช้ในการเลือกคำสำคัญที่เป็นที่นิยมมีดังนี้

ตาราง 2.1 ตัวอย่างฟังก์ชันการคำนวณหาค่าน้ำหนักสำหรับการเลือกลักษณะเด่น

Function	Denoted by	Mathematical form
DIA association factor	$z(t_k, c_i)$	$P(c_i \mid t_k)$
Information gain	$IG(t_k, c_i)$	$\sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t, c). log \frac{P(t, c)}{P(t). P(c)}$
Mutual information	$MI(t_k, c_i)$	$log \frac{P(t_k, c_i)}{P(t_k). P(c_i)}$
Chi-square	$x^2(t_k, c_i)$	$\frac{\lvert Tr \rvert . [P(t_k, c_i). P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i). P(\bar{t}_k, c_i)]^2}{\sqrt{P(t_k). P(\bar{t}_k). P(c_i). P(\bar{c}_i)}}$
NGL coefficient	$NGL(t_k, c_i)$	$\frac{\sqrt{\lvert Tr \rvert} . [P(t_k, c_i). P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i). P(\bar{t}_k, c_i)]}{\sqrt{P(t_k). P(\bar{t}_k). P(c_i). P(\bar{c}_i)}}$
Relevancy score	$RS(t_k, c_i)$	$log \frac{P(t_k \mid c_i) + d}{P(\bar{t}_k \mid \bar{c}_i) + d}$
Odds ratio	$OR(t_k, c_i)$	$\frac{P(t_k \mid c_i). (1 - P(t_k \mid \bar{c}_i))}{(1 - P(t_k \mid c_i)). P(t_k \mid \bar{c}_i)}$
GSS coefficient	$GSS(t_k, c_i)$	$P(t_k, c_i). P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i). P(\bar{t}_k, c_i)$

(Sebastiani, 2002)

จากตาราง 2.1 ใน Mathematical form โดยค่าที่คำนวณได้ขึ้นอยู่กับเหตุการณ์ของเอกสาร เช่น $P(F_k, c_i)$ คือความเป็นไปได้ของการสุ่มเอกสาร x โดยที่คำสำคัญ t_k ไม่ได้ปรากฏในเอกสาร x และ x คือเอกสารที่เป็นสมาชิกของกลุ่ม c_i

แต่ละฟังก์ชัน $f(t_k, c_i)$ การกำหนดกลุ่ม c_i และคำสำคัญ t_k ในการหาค่าของแต่ละฟังก์ชัน ในการหาค่าน้ำหนักแต่ละคำสำคัญได้จาก การหาค่าเฉลี่ย (Average) $f_{avg} = \sum_{i=1}^{m}(t_k, c_i) . P(c_i)$ การหาค่าสูงสุด (Maximum) $f_{max}(t_k) = \max_{i=1}^{m} f(t_k, c_i)$ หรือการหาผลรวมของค่าน้ำหนัก (Summation) $f_{sum}(t_k) = \sum_{i=1}^{|c|} P(c_i)(t_k, c_i)$

2.3.2 การลดมิติโดยการสกัดคำสำคัญ (Dimensionality reduction by term extraction)

กระบวนการสกัดคำสำคัญเรียกอีกอย่างว่าการสกัดลักษณะเด่น เป็นกระบวนการเปลี่ยนการแทนเอกสารแบบเดิมในรูปของดัชนีเอกสาร ให้อยู่ในรูปแบบของการแทนเอกสารแบบอื่น ซึ่งอยู่บนพื้นฐานของการลดมิติของเอกสาร และยังคงรักษาเนื้อหาหรือความแตกต่างของเอกสารได้ การสกัดลักษณะเด่นที่นิยมใช้ และมีประสิทธิภาพสูง คือ การหาโครงสร้างความหมายภายในประกอบด้วยกระบวนการดังนี้

1) กระบวนการสร้างดัชนีความหมายภายใน

กระบวนการดัชนีความหมายภายใน เรียกอีกอย่างว่าแอลเอสไอ คือ กระบานกานสกัดลักษณะเด่น ที่เป็นที่นิยมและมีประสิทธิภาพสูง โดยใช้เทคนิคในการสร้างตัวแทนของคิวรี (Query) และเอกสารให้อยู่ในรูปของพื้นที่ความหมายภายใน (Latent semantic space) ซึ่งตัวแทนเอกสารนี้จะมีจำนวนมิติที่น้อยลงกว่ามิติที่เอกสารที่ถูกแทนด้วยเมตริกซ์คำสำคัญ – เอกสาร อย่างไรก็ตาม แอลเอสไอ ถูกสร้างขึ้นเพื่องานด้านการค้นคืนเอกสารและแอลเอสไอยังถูกนำไปใช้ในงานจำแนกเอกสารด้วย โดยลักษณะเด่นของแอลเอสไอคือ การลดมิติของข้อมูล และการแก้ปัญหาด้านความหมายของคำ เช่น คำที่เขียนต่างกันแต่ความหมายเหมือนกัน (Synonym) และคำที่มีความหมายต่างกัน แต่เขียนเหมือนกัน (Polysemy)

กระบวนการทำงานของแอลเอสไอได้มีการนำกระบวนการทางคณิตศาสตร์ที่เรียกว่าเอสวีดี (Singular values decomposition: SVD) ในการแปลงเมตริกซ์คำสำคัญ-เอกสาร โดยการสร้างพื้นที่ความหมายภายใน เพื่อใช้ในการสร้างตัวแทนเอกสารให้อยู่ในรูปของ เค-มิติ (k-dimension) โดยที่ k มีขนาดเล็กกว่าจำนวนของคำสำคัญทั้งหมดมาก และพยายามรักษาเนื้อหาเดิมของเอกสารให้มากที่สุด

การคำนวณแอลเอสไอโดย เอสวีดี คือการสร้างเมตริกซ์ตัวแทน A′ จากเมตริกซ์ A ซึ่ง A′ คือการประมาณค่าของเมตริกซ์ A จาก เค-มิติที่เลือกมา เมื่อทำการคำนวณเอสวีดีบนเมตริกซ์คำสำคัญ-เอกสาร A_{txd} ผลลัพธ์ที่ได้ประกอบด้วย T_{txn} , S_{nxn} และ D_{dxn} โดยมีสมการดังนี้

$$A_{txd} = T_{txn} \times S_{nxn} \times (D_{dxn})^T$$

โดยที่ t คือจำนวนคำสำคัญ d คือจำนวนเอกสารและ n คือค่าต่ำสุดระหว่าง t และ d , T คือ ค่าออร์-โทโกนอล (Orthogonal values) เช่น $TT^T = D^TD = I$, S คือ Diagonal($\sigma_1, \sigma_2, ..., \sigma_n$) โดยที่ $\sigma_i > 0$

T_{txn} คือ เมตริกซ์ของคอลัมน์เป็นไอเกนเวคเตอร์ (Eigenvector) ของ AA^T หรือเรียกว่าไอเกนเวคเตอร์ซ้าย (Left eigenvector)

S_{nxn} คือ เมตริกซ์ของไดอะโกนอล (Diagonal) คือค่าซิงกูลาร์ (Singular values) ของ A

D_{dxn} คือ เมตริกซ์ของคอลัมน์เป็นไอเกนเวคเตอร์ของ A^TA หรือเรียกอีกอย่างว่าไอเกนเวคเตอร์ขวา (Right eigenvector)

จะเห็นได้ว่าเอสวีดี คือกระบวนการสำหรับการหมุนแกนมิติ โดยมิติแรกแสดงความแตกต่างกันมากที่สุดระหว่างเอกสาร มิติที่สองแสดงความแตกต่างของเอกสารมากที่สุดลำดับที่สองและมิติถัดไปก็จะแสดงความแตกต่างมากที่สุดลำดับถัดไปเช่นกัน

ในการเลือกมิติที่ 1 ถึง k มิติโดยที่ k น้อยกว่า n จากเมตริกซ์ T, S และ D ในการประมาณค่าของ A ได้ว่า

$$A'_{txd} = T_{txk} \times S_{kxk} \times (D_{dxk})^T$$

1.1) การสร้างตัวแทนดิวรีหรือการสร้างตัวแทนเอกสารในรูปแบบของแอลเอสไอ

สำหรับงานทางด้านการค้นคืนเอกสารและงานด้านการจำแนก ดิวรี หรือเอกสาร ซึ่งทั้งคู่อยู่ในรูปของกลุ่มของคำสำคัญ ต้องถูกแปลงให้อยู่ในรูปของ เค-มิติ ในรูปแบบของเอสวีดี ในการเปรียบเทียบหรือในการจำแนกกลุ่ม โดยการสร้างตัวแทนมีขั้นตอนดังนี้

$$q' = q^T . T_{txk} . S_{kxk}^{-1}$$

โดยที่ q คือดิวรีหรือเอกสาร อยู่ในรูปของเวคเตอร์ของคำสำคัญ (Barbara, 2000)

1.2) การหาค่าเอสวีดี

เมื่อคำนวณเอสวีดี บนเมตริกซ์ A โดยที่ A มีขนาด txd ดังสมการ

$$A_{txd} = T_{txn} \times S_{nxn} \times (D_{dxn})^T$$

โดยที่ t คือจำนวนคำสำคัญ , d คือจำนวนเอกสาร , n คือค่าต่ำสุดระหว่าง d และ n

โดยที่ $T^T T = I_{nxn}$,$D^T D = I_{nxn}$

ซึ่ง T และ D คือ ออร์โทโกนอล

โดยที่ T คือ ไอเกนเวคเตอร์, S คือ ค่าซิงกูลาร์, D คือ ไอเกนเวคเตอร์ขวา

การคำนวณ เอสวีดี ขึ้นอยู่กับการคำนวณหาค่าของไอเกนเวคเตอร์และค่าไอเกนของ AA^T และ A^TA โดยที่ไอเกนเวคเตอร์ของ A^TA ใช้ในการสร้างคอลัมน์ของ D ส่วนไอเกนเวคเตอร์ของ AA^T ใช้ในการสร้างคอลัมน์ของ T และ S สร้างจากรากที่สองของ AA^T หรือ A^TA โดยค่าซิงกูลาร์เป็นการเรียงลำดับจากมากไปน้อยของค่าไดอะโกนอล ของเมตริกซ์ S ซึ่งค่าซิงกูลาร์คือ จำนวนจริง (Real number) ถ้า A คือจำนวนจริงจะทำให้ T และ D เป็นจำนวนจริงไปด้วย ในการหาค่าไอเกนของ A ได้จาก

$$W = A^TA = AA^T$$

$$Wx = \lambda x$$

สำหรับบางปริมาณของ λ ซึ่งปริมาณของ λ คือค่าไอเกนของ A และ x คือ ไอเกนเวคเตอร์ของ A จะได้ว่า

$$(W - \lambda I)x = 0$$

โดยที่ I คือ เมตริกซ์เอกลักษณ์

ในการค่าหา S หาได้จากปริมาณของ λ จากนั้นทำการเรียงลำดับค่า λ จากมากไปน้อย

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$$

ต่อไปทำการหารากที่สองของปริมาณ λ

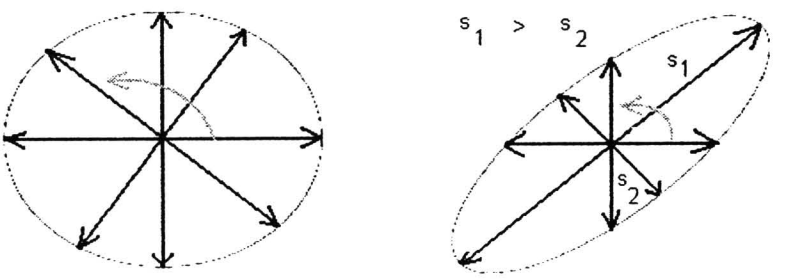
$$S_i = \sqrt{|\lambda_i|}$$

และทำการสร้างเมตริกซ์ โดยใช้ S_i เรียงลำดับในตำแหน่งไดอโกนอล์

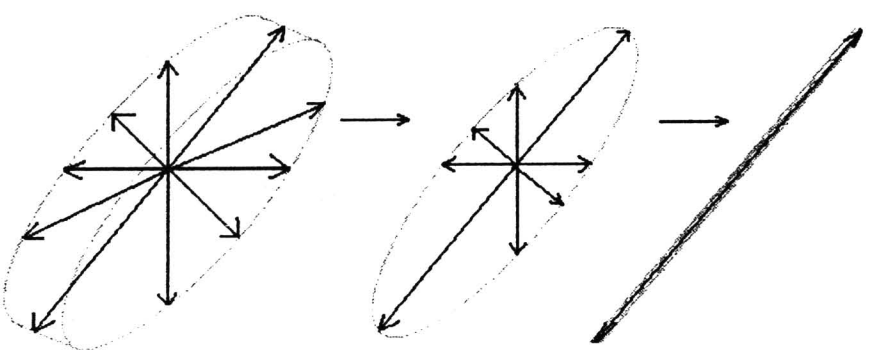
$$\begin{bmatrix} S_1 & 0 & \cdots \\ 0 & S_2 & \vdots \\ \vdots & \vdots & \vdots \\ \cdots & \cdots & S_n \end{bmatrix}$$

1.3) การกำหนดค่า T, D

การหาค่า T เมื่อเราคำนวณหาค่าของ ค่าไอเกนให้ทำการคำนวณหาค่าไอเกนเวคเตอร์ x จากค่าไอเกนแต่ละค่าทีคำนวณได้จาก $(W - \lambda_i I)x_i = 0$ โดยที่ $W = A.A^T$ จากนั้นทำการสร้างเมตริกซ์จากค่าที่คำนวณ ซึ่งในการหาค่าของเมตริกซ์ D คำนวณตามขั้นตอนของ T โดยที่ $W = A^T.A$



รูป 2.2 แสดงการหมุนแกน



รูป 2.3 แสดงการลดมิติ

(Garcia, 2005)

2) กระบวนการพีแอลเอส (Partial latent square : PLS)

กระบวนการแอลเอสไอ มีพื้นฐานมาจากการสร้างโครงสร้างความหมายภายในจากเมตริกซ์คำสำคัญ – เอกสาร จากคำสำคัญที่มีจำนวนมาก อาจจะส่งผลให้ประสิทธิภาพการสร้างตัวแทนเอกสารลดลง ปัญหาหนึ่งในการใช้แอลเอสไอในการสร้างตัวแทนเอกสาร คือลักษณะสำคัญของบางกลุ่มที่มีขนาดเล็กจะไม่ถูกนำมาพิจารณา ซึ่งในความเป็นจริงคือ บางกลุ่ม มักจะแสดงลักษณะสำคัญออกมาจากคำสำคัญบางคำแต่คำสำคัญนั้นๆ ไม่ถูกสร้างความหมายลงในพื้นที่ความหมายภายใน ในการสร้างตัวแทนเอกสาร และคำสำคัญนั้นๆอาจจะถูกพิจารณาว่าเป็นคำที่ส่งผลให้ความแตกต่างระหว่างกลุ่มผิดเพี้ยนไป การแก้ไขปัญหานี้โดยใช้โลคอลแอลเอสไอ (Local LSI) แทนแอลเอสไอ โดยการสร้างแอลเอสไอของแต่ละกลุ่มเอกสาร (Wiener, 1995) อย่างไรก็ตามโลคอลแอลเอสไอ มีข้อเสียคือ

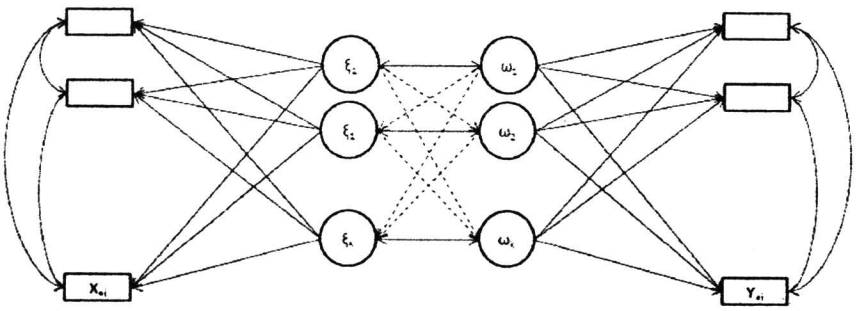
(1) เมื่อมีเอกสารที่ต้องจำแนกกลุ่มเอกสารนั้นจะถูกสร้างตัวแทนเอกสารในแต่ละกลุ่ม ด้วยแอลเอสไอ จึงส่งผลให้ใช้เวลาในการคำนวณสูง และเอกสารนั้นๆ อาจถูกพิจารณาว่ามีความคล้ายกับกลุ่มใดกลุ่มหนึ่งมาก ทั้งๆที่เอกสารนั้นไม่ได้เป็นสมาชิกของกลุ่มนั้นเลย

(2) กระบวนการโลคอลแอลเอสไอ ไม่สามารถแก้ปัญหาคำที่คลุมเครือได้ จึงอาจทำให้การจำแนกกลุ่มลดประสิทธิภาพลง

ในการแก้ปัญหาโดย Wang (Wang, and Nie, 2003) ได้มีการเสนอแนวทางการแก้ปัญหา โดยการพิจารณา เมตริกซ์เอกสาร – คำสำคัญ และ เมตริกซ์เอกสาร – กลุ่มของเอกสารในเวลาเดียวกัน กระบวนการนี้ คล้ายกับแอลเอสไอ ในการวางคำสำคัญลงบนพื้นที่ความหมายภายใน แต่กระบวนการนี้ สร้างความสัมพันธ์ของความหมายภายใน (Latent semantic relationship) ระหว่างคำสำคัญและกลุ่มเอกสาร และรักษาเนื้อหาของเอกสารไปพร้อมๆกัน ซึ่งถ้ามองในมุมมองทางสถิติจะพบว่าคล้ายกับกระบวนการวิเคราะห์พีแอลเอส (Partial latent square analysis) โดยมีข้อกำหนดดังนี้ X คือตัวแทนของเมตริกซ์เอกสาร – คำสำคัญ m x n และ Y คือเมตริกซ์เอกสาร – กลุ่มของเอกสาร m x r

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1r} \\ \vdots & \vdots & \cdots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mr} \end{pmatrix}$$

โดยที่ y_{ij} มีค่าเป็น 1 เมื่อเอกสาร i อยู่ในกลุ่ม j และมีค่าเป็น 0 ในกรณีอื่นๆ สามารถมองความสัมพันธ์ระหว่าง X และ Y ดังภาพต่อไปนี้



รูป 2.4 แสดงรูปแบบความสัมพันธ์ของโครงสร้างภายในของกระบวนการพีแอลเอส

ในรูป 2.4 รูป ◯ คือตัวแปรภายใน (Latent variable) ξ และ ω รูป ▭ คือตัวแปรคำสำคัญ X_q และกลุ่ม Y_q หัวลูกศรคู่ระหว่าง ξ และ ω แสดงดังค่าสหสัมพันธ์ (Correlation) ที่มากกว่า 0

- หัวลูกศรเดี่ยวระหว่างตัวแปรภายในชี้ไปยัง คำสำคัญหรือกลุ่มของข้อมูล คือค่าสัมประสิทธิ์ (Coefficient) ที่มากกว่า 0
- ไม่มีเส้นเชื่อม คือความขึ้นต่อกันสามารถเกิดได้จากการที่ตัวแปรอื่นมีการส่งผ่านความสัมพันธ์
- หัวลูกศรคู่เส้นหนา ระหว่างตัวแปรคำสำคัญ คือระหว่างตัวแปรคำสำคัญมีความเกี่ยวเนื่องกันในการให้ค่า ξ ซึ่งสามารถเกิดกับตัวแปรกลุ่มในการให้ค่า ω เช่นกัน
- ลูกศรที่เป็นเส้นประ คือระหว่างตัวแปรมีความสัมพันธ์กันจริง แต่มีความสัมพันธ์กันไม่ชัดเจน

จากรูป 2.4 พบว่าไม่ได้มีการนำค่าความแปรปรวนร่วม (Covariance) ของตัวแปรคำสำคัญของเมตริกซ์ของ X มาพิจารณา แต่กลับสนใจค่าความแปรปรวนร่วมแบบไขว้ (Cross – covariance) ระหว่าง X และ Y และต้องการสร้างค่าความแปรปรวนร่วมแบบไขว้ระหว่างคู่ตัวแปรภายใน เช่น $(\xi_1, \omega_1), (\xi_2, \omega_2),\ldots,(\xi_k, \omega_k)$

โดย ξ_i แทน เนื้อหาความหมายภายใน (Latent semantic information) ของ X และ ω_i แทน เนื้อหาความหมายภายในของ Y และ (ξ_i, ω_i) คือการลดจำนวนของค่าที่สำคัญของเนื้อหาความหมายภายในของเมตริกซ์ X และ Y ซึ่งมีพื้นฐานในการพิจารณา (ξ_i, ω_i) ดังนี้

- (ξ_1, ω_1) คือคู่ของตัวแปรที่ดีที่สุดระหว่าง X และ Y
- (ξ_2, ω_2) คือคู่ของตัวแปรที่ดีที่สุดระหว่าง X และ Y ลำดับที่สอง
- (ξ_k, ω_k) คือคู่ของตัวแปรที่ดีที่สุดระหว่าง X และ Y ที่เหลือจากการเกิดค่าของลำดับก่อนหน้า

โดยคู่แรกของตัวแปรภายใน (ξ_1, ω_1) มีคือกำหนดดังนี้

1) ξ_1 แทนเนื้อหาความหมายภายในของ X ที่ดีที่สุดที่เป็นไปได้ ในทางสถิติคือค่าแปรปรวน (Variance) สูงสุด $Var(\xi_1) \rightarrow max$

2) ω_1 แทนเนื้อหาความหมายภายในของ Y ที่ดีที่สุดที่เป็นไปได้ ในทางสถิติคือแปรปรวน สูงสุด $Var(\omega_1) \rightarrow max$

3) (ξ_1, ω_1) ต้องแทนค่าสหสัมพันธ์ ระหว่าง X และ Y ที่ดีที่สุดที่เป็นไปได้ ในทางสถิติคือค่าสัมประสิทธิสหสัมพันธ์ (Correlation coefficient) สูงสุด $r(\xi_1, \omega_1) \rightarrow max$

ถ้าพิจารณา ξ_1 คือส่วนประกอบของคำสำคัญ สามารถแสดงในรูปสมการดังนี้

$$\xi_1 = Xu$$

โดยที่ u คือเวคเตอร์ที่ถูกพิจารณา ξ_1 คือ ลิเนียร์ คอมไบเนชัน (Linear combination) ของคำสำคัญบางคำในเอกสาร ซึ่งมีลักษณะเช่นเดียวกันกับหน่วยความหมายที่สำคัญและสามารถแสดง ω_1 ดังสมการ

$$\omega_1 = Yv$$

โดยที่ v คือเวคเตอร์ที่ถูกพิจารณา ω_1 คือ ลิเนียร์ คอมไบเนชันของกลุ่มข้อมูลความสัมพันธ์ภายในระหว่างคำสำคัญของเอกสารและกลุ่มของเอกสาร สามารถแปลงเงื่อนไขออกมาเป็นสมการได้ดังนี้

$$\max(\mathrm{var}(\xi_1)) = \max_{\|u\|=1}(\mathrm{var}(Xu))$$

$$\max(\mathrm{var}(\omega_1)) = \max_{\|v\|=1}(\mathrm{var}(Yv))$$

$$\max(r(\xi_1,\omega_1)) = \max_{\|u\|=\|v\|=1} r(Xu, Yv)$$

โดยที่ ||u|| และ ||v|| แทนขนาดของเวคเตอร์ u และ v

ดังนั้นสามารถหาค่าความแปรปรวนร่วมได้จากสมการ

$$\mathrm{Cov}(\xi_1,\omega_1) = \sqrt{\mathrm{var}(\xi_1)\,\mathrm{var}(\omega_1)} \times r(X_u, Y_v)$$

ซึ่งการหาค่าสูงสุดได้ว่า

$$\mathrm{Cov}(\xi_1,\omega_1) \rightarrow max$$

ให้ (.,.) แทนดอทโปรดักท์ (Dot Product) เนื่องจาก $(\xi_1,\omega_1) = \mathrm{Cov}(\xi_1,\omega_1)$ ในการพิจารณา (ξ_1,ω_1) สามารถแปลงให้อยู่ในปัญหาค่าสูงสุด (Maximization problem)

$$(\xi_1,\omega_1) = \xi_1^T\omega_1 = \max_{\|u\|=\|v\|=1}(Xu, Yv)$$

ให้ u_1 และ v_1 คือ โซลูชันเวคเตอร์ (Solution vector) ของปัญหาค่าสูงสุดและจากปัญหาของเอสวีดี $d_1u_1v_1^T$ คือค่าที่ดีที่สุดในมิติแรกของ X^TY โดยที่ d_1 คือค่าซิงกูลาร์ใหญ่ที่สุด (Largest singular values) u_1และ v_1 คือค่าซิงกูลาร์เวคเตอร์ซ้ายและขวา

ตัวแปรภายในคู่แรก (ξ_1, ω_1) ถูกสร้างขึ้นเพื่อให้สามารถจัดเก็บการแทนเนื้อหา โดยการประมาณค่า x บน ξ_1 และประมาณค่า y บน ω_1 ดังต่อไปนี้

$$\hat{X}_1 = \xi_1(\xi_1^T \xi_1)^{-1}\xi_1^T X$$

$$\hat{Y}_1 = \omega_1(\omega_1^T \omega_1)^{-1}\omega_1^T Y$$

จากนั้นทำการหาค่าที่เหลือจาก $X = X - \hat{X}$ และ $Y = Y - \hat{X}$ ขั้นตอนต่อไป นำค่า X และ Y ที่ได้ ไปใช้ในการหาค่า (ξ_2, ω_2) และคู่อื่นต่อไป ด้วยเหตุนี้จึงมีการใช้กระบวนการวิธีหาพีแอลเอสมาใช้ในกระบวนการลดมิติของเอกสาร โดยขั้นตอนวิธีพีแอลเอสมีขั้นตอนดังนี้

ALGORITHM-1(LSR/PLS2)

$E_0 = X;$ $F_0 = Y;$
FOR k = 1, ..., s DO
  $u_k =$ first column of F_{k-1};
  Do until convergence in ξ_k// computing pairs of latent variables
    $\xi_k = E_{k-1}^T u_k / u_k^T u_k$
    $\xi_k = \frac{\xi_k}{||\xi_k||}$
    $t_k = E_{k-1}\xi_k$
    $\omega_k = F_{k-1}^T t_k / t_k^T t_k$
    $u_k = F_{k-1}\omega_k / \omega_k^T \omega_k$
  ENDDO
$P_k = F_{k-1}^T t_k / t_k^T t_k$
$E_k = E_{k-1} t_k P_k^T$ //remaining information of document - term matrix
$F_k = F_{k-1} t_k \omega_k^T$ //remaining information of document - class matrix
ENDFOR (Wang and Nie, 2003)

เมื่อได้เมตริกซ์ ξ ทำการสร้างตัวแทนเอกสารดังนี้

$$q' = q \times \xi$$

โดยที่ q' คือตัวแทนเอกสารของเอกสาร q (Zeng, Wang and Nie, 2007)

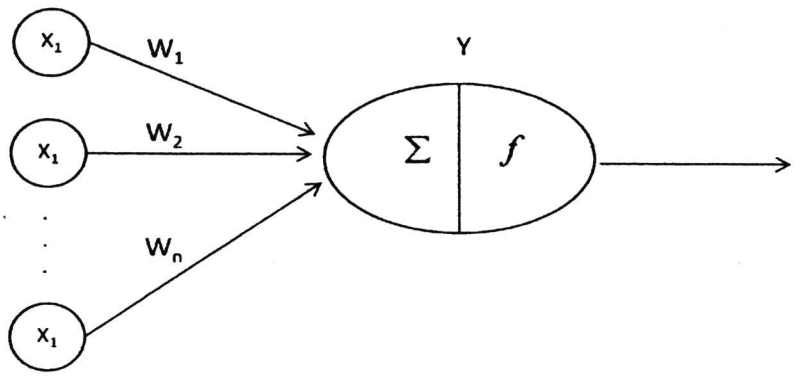
2.4 ทฤษฎีโครงข่ายประสาท

2.4.1 โครงสร้างพื้นฐานของโครงข่ายประสาท

โครงข่ายประสาทมีแนวคิดที่เลียนแบบการทำงานของระบบประสาทของมนุษย์ โดยพยายามพัฒนาคอมพิวเตอร์ให้มีการเรียนรู้และสามารถตัดสินใจได้เอง โดยอาศัยแบบจำลองทางคณิตศาสตร์เข้ามาช่วยในการคำนวณ

โครงข่ายประสาทจะประกอบไปด้วยหน่วยประมวลผล เรียกว่าเซลล์ประสาท (Neuron) ที่เชื่อมถึงกันหลายๆตัว ผ่านการเชื่อมต่อที่เรียกว่าเส้นเชื่อม (Edge) แต่ละเส้นเชื่อม จะมีค่าน้ำหนัก

(Weight) กำกับอยู่ แต่ละเซลล์ประสาท จะมีค่าสถานะภายในที่เรียกว่าระดับการกระตุ้น (Activity level) ข้อมูลออกจาก (Output) เซลล์ประสาท แรกที่ส่งไปยังเซลล์ประสาท ถัดไปจะได้มาจากการพิจารณาผลรวมทั้งหมดของค่าน้ำหนักคูณกับค่าของข้อมูลเข้า (Input) ดังแสดงในรูปที่ 2.5



รูป 2.5 แสดงเซลล์ประสาทของโครงข่ายประสาท

จากรูป 2.5 ข้อมูลเข้า คือ $X_1, X_2,..., X_n$ ถูกป้อนให้กับเซลล์ประสาท Y ข้อมูลเข้านี้จะถูกนำไปคูณเข้ากับค่าน้ำหนักคือ $W_1, W_2,..., W_n$ จากนั้นทำการหาผลรวมค่าที่ได้ ต่อไปจะมีการพิจารณาข้อมูลออก โดยอาศัยฟังก์ชันกระตุ้น (Activation faction)

ค่าผลรวมของข้อมูลเข้าคูณกับน้ำหนักได้จากสมการ

$$Y = \sum_{i=0}^{n} X_i W_i + \theta$$

โดย θ คือ ค่าเบี่ยงเบน (Bias)

ค่าสัญญาณเซลล์ประสาทออก Y ได้จากสมการ

$$Out = f(Y)$$

โดยที่ f คือฟังก์ชันกระตุ้น ซึ่งกำหนดได้ดังนี้

$$f(Y) = \frac{1}{1 + e^{-Y}}$$

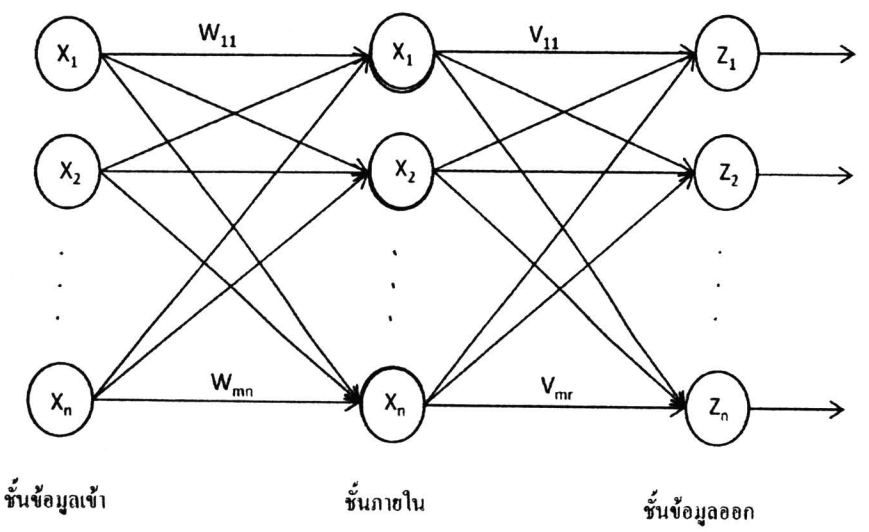
(Hu and Hwang, 2001)

2.4.2 โครงข่ายประสาทชนิดการเรียนรู้แบบแบคพรอพาเกชัน (Backpropagation)

หลักการการเรียนรู้แบบแบคพรอพาเกชันเป็นขั้นตอนที่ใช้ในการเรียนรู้โครงข่ายประสาทแบบหลายชั้น โดยขั้นตอนในการปรับค่าน้ำหนักเพื่อให้ได้ค่าที่เหมาะสมนั้นจะใช้วิธีเรียนรู้เป้าหมายของแต่ละข้อมูลเข้าคืออะไร และใช้ความผิดพลาดของข้อมูลออกเปรียบเทียบกับ

เป้าหมาย เพื่อเป็นตัวชี้นำในการปรับน้ำหนัก สำหรับชั้นภายในจะไม่มีค่าเป้าหมายที่จะทำการเปรียบเทียบ ดังนั้นการปรับค่าน้ำหนักสำหรับชั้นภายในจึงใช้วิธีการแพร่ค่าความผิดพลาดจากชั้นข้อมูลออก กลับมายังชั้นภายใน กระบวนการสำคัญของการเรียนรู้แบบแพร่กลับมี 3 ขั้นตอน คือ การป้อนไปข้างหน้า (Feed forward) ของรูปแบบข้อมูลเข้า (Input pattern) การคำนวณและส่งค่าผิดพลาดกลับคืน (Back propagation of Error) และการปรับค่าน้ำหนักให้เหมาะสม

โครงข่ายประสาทชนิดการเรียนรู้แบบแบคพรอพาเกชันประกอบด้วย 3 ชั้น คือ ชั้นข้อมูลเข้า ชั้นภายใน และชั้นข้อมูลภายนอก ดังรูป 2.6

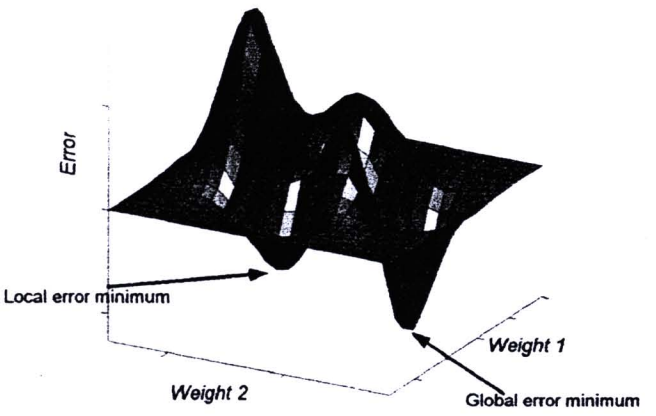


รูป 2.6 แสดงโครงข่ายประสาทชนิดการเรียนรู้แบบแพร่กลับ

กระบวนการการทำงานของแบคพรอพาเกชัน เริ่มจากการกำหนดค่าน้ำหนักให้แต่ละเส้นเชื่อมในโครงข่ายประสาท โดยการสุ่มค่าขนาดเล็กมีค่าระหว่าง 0 ถึง 1 จากนั้นทำการรับข้อมูลเข้าให้แก่โครงข่าย เพื่อให้โครงข่ายทำการทำนายผลลัพธ์ จากนั้นทำการคำนวณค่าความผิดพลาดระหว่างผลลัพธ์จากการทำนาย เปรียบเทียบกับผลลัพธ์ที่แท้จริง ดังต่อไปนี้

$$\sum_{p=1}^{P}\sum_{i=1}^{S}(t_i^p - o_i^p)^2$$

เมื่อ P คือ จำนวนเส้นเชื่อมของข้อมูลเข้า S คือ คำนวณผลลัพธ์ของโครงข่ายประสาท t_i^p คือ ค่าที่แท้จริง และ o_i^p คือค่าที่ได้จากการทำนาย



รูป 2.7 แสดงพื้นผิวของความผิดพลาด (Error surface) จากการปรับค่าน้ำหนักสองครั้ง

แบคพรอพาเกชัน ทำการลดความผิดพลาด ด้วยการคำนวณจากกราเดียท์ (Gradient) ของพื้นผิวของความผิดพลาด การปรับค่าน้ำหนักของโครงข่าย โดยการส่งค่าความผิดพลาดย้อนกลับไป เริ่มจาก ชั้นข้อมูลออกไปถึงชั้นข้อมูลเข้า การปรับค่าน้ำหนักจะพยายามให้ได้ค่าการทำนายที่ดีที่สุดในระดับที่ยอมรับได้ เมื่อเปรียบเทียบกับผลลัพธ์ที่แท้จริง ในการหยุดการเรียนรู้ของโครงข่ายจะพิจารณาจากค่าผิดพลาดต่ำสุด (Minimum error) โดยมีขั้นตอนพื้นฐานดังนี้

1) กำหนดค่าเริ่มต้นของค่าน้ำหนักโดยการสุ่มค่าที่มีขนาดเล็ก โดยมีค่าระหว่าง 0 ถึง 1

2) ทำการป้อนข้อมูลเข้า และ ค่าที่แท้จริง

3) ทำการคำนวณค่าความผิดพลาดจากการเปรียบเทียบผลลัพธ์จากการทำนาย และค่าที่แท้จริง

4) การปรับค่าน้ำหนักโดยการกระจายความผิดพลาดไปให้แก่เส้นเชื่อมแต่ละเส้นจากชั้นข้อมูลออกไปจนถึงชั้นข้อมูลเข้า

5) ทำการทำซ้ำกระบวนการที่ 2) ถึง 4) จนกว่าค่าความผิดพลาดไม่มีการเปลี่ยนแปลงหรือเท่ากับค่าใดค่าหนึ่งที่กำหนดไว้(Anthony and Michael, 2006)

การสร้างแบบจำลองโครงข่ายประสาทแบบแบคพรอพาเกชันมีขั้นตอนวิธีดังนี้

1) กำหนดค่าเริ่มต้นสำหรับค่าน้ำหนัก w กับค่าเบี่ยงเบน θ โดยใช้การสุ่มค่าโดยเลือกค่าน้อยๆ

2) คำนวณค่าของข้อมูลเข้า I_j กับค่าข้อมูลออก O_j

(1) คำนวณค่า I_j

$$I_j = \sum_{i=1}^{n} w_{ij} o_i + \theta_j$$

โดยที่ w_{ij} = ค่าน้ำหนักของเส้นเชื่อมระหว่างหน่วย i ในชั้นก่อนหน้าถึงหน่วย j

o_i = ข้อมูลออกของหน่วย i จากชั้นก่อนหน้า

θ_j = ค่าเบี่ยงเบน

(2) คำนวณค่า O_j

$$O_j = \frac{1}{1 + e^{-Ij}}$$

3) คำนวณค่าผิดพลาดย้อนกลับสำหรับ O_j และ Err_j

(1) ความผิดพลาดจากหน่วยของชั้นข้อมูลออก j

$$Err_j = O_j(1 - O_j)(T_j - O_j)$$

โดยที่ o_i = ผลที่ได้จากหน่วย j

T_j = ผลลัพธ์ที่แท้จริง (มาจากกลุ่มที่กำหนดไว้แล้ว)

(2) ความผิดพลาดจากหน่วยของชั้นข้อมูลออก j

$$Err_j = O_j(1 - O_j) \sum_k Err_k w_{jk}$$

โดยที่ w_{jk} = ค่าน้ำหนักของเส้นเชื่อมระหว่างหน่วย j ถึงหน่วย k ในชั้นถัดไป (Next higher layer)

Err_k = ความผิดพลาดจากหน่วย k

$\sum_k Err_k w_{jk}$ = ผลรวมของค่าน้ำหนักของความผิดพลาดระหว่างหน่วย j กับชั้นถัดไป

(3) ทำการปรับค่าน้ำหนัก w กับค่าเบี่ยงเบน θ เพื่อลดความผิดพลาดให้มีค่าต่ำที่สุด

$$\Delta w_{ij} = (1) Err_j O_j$$
$$w_{ij} = w_{ij} + \Delta w_{ij}$$
$$\Delta \theta_j = (1) Err_j$$
$$\theta_j = \theta_j + \Delta \theta_j$$

โดยที่ Δw_{ij} = ค่าน้ำหนักที่เปลี่ยนแปลง

$\Delta \theta_j$ = ค่าเบี่ยงเบนที่เปลี่ยนแปลง

1 = ระดับการเรียนรู้ (มีค่าระหว่าง 0.0 ถึง 1.0)

(Han and Kamber, 2001)