



UNIVERSITÉ  
JEAN MONNET  
SPÉCIALITÉ WEB  
INTELLIGENCE

TSINGHUQ  
UNIVERSITY  
DEPARTMENT OF  
ELECTRONIC ENGINEERING

STAGE EN ENTREPRISE: STAGE DE FIN D'ÉTUDE

---

# Rapport de Stage Année 2013-2014

Fouille de Donnée dans la domaine de  
télécommunication

---

*Auteur :*  
Wenyi WANG

*Tuteur de stage en entreprise :*  
Vice directeur de labo NGN : yongfeng  
HUANG

*Tuteur de l'université :*  
Amaury HABRARD

De 20 Février 2014 à 20 Juillet 2014

# Table des matières

<b>Remerciements</b>	<b>3</b>
<b>1 Résumé</b>	<b>1</b>
<b>2 Introduction</b>	<b>1</b>
2.1 Introduction du CMCC . . . . .	1
2.2 La crise du CMCC . . . . .	2
2.2.1 La demande de mise à niveau du réseau . . . . .	2
2.2.2 Les changements des moyennes de revenu . . . . .	4
2.3 L'optimisation du réseau . . . . .	5
2.4 Introduction du laboratoire . . . . .	6
2.5 Objectif du projet . . . . .	6
<b>3 Introduction de l'industrie de la télécommunication</b>	<b>7</b>
3.1 L'évolution des normes de téléphonie mobile . . . . .	7
3.1.1 La premier génération . . . . .	8
3.1.2 La deuxième génération . . . . .	8
3.1.3 La troisième génération . . . . .	9
3.1.4 La quatrième génération . . . . .	9
3.2 Le réseau LTE . . . . .	10
3.2.1 La structure du réseau LTE . . . . .	11
<b>4 Les solution existant</b>	<b>13</b>
<b>5 Le présentation de notre solution</b>	<b>16</b>
5.1 Clustering . . . . .	17
5.1.1 La distance . . . . .	18
5.1.2 La validité de clustering . . . . .	19
5.1.3 La silhouette Coefficient . . . . .	20
5.2 Règles d'association . . . . .	21

5.3	K-Means et l'arbre couvrant de poids minimal . . . . .	22
5.3.1	L'algorithme KmMST . . . . .	23
<b>6</b>	<b>La mise en œuvre</b>	<b>24</b>
6.1	Le logiciel utilisés . . . . .	24
6.2	Introduction des données . . . . .	25
6.3	Prétraitement de données . . . . .	26
6.4	Les caractéristique du donnée . . . . .	29
6.5	Le K-Means et Le Règle d'association . . . . .	30
6.5.1	La valeur optimal de $k$ . . . . .	30
6.5.2	Le clustering . . . . .	31
6.5.3	Le Règle d'association . . . . .	33
6.6	Le KmMST . . . . .	35
6.6.1	L'étape de l'algorithme KmMST . . . . .	35
	<b>Conclusion</b>	<b>38</b>
	<b>Références</b>	<b>40</b>

## Remerciements

Tout d'abord, je tiens à remercier Amaury Habrard et tous les enseignants de la Spécialité Web Intelligence de l'Université Jean Monnet, aussi les enseignants de Télécom Saint-Etienne et L'école nationale supérieure de Saint-Etienne, qui m'a aidé lors de ces deux années de étude.

Je remercie également M.Yongfeng HUANG pour avoir accepté diriger cette stage, il m'a beaucoup conseillé, et les discussions que l'on a pu avoir se sont toujours révélées très intéressantes et instructives.

Je souhaite également adresser mes remerciement à Zheng YANG, Lindong WEI et xian WU ainsi que tout les membres du laboratoire de Next generation Network( **NGN** ) pour m'avoir soutenu, encouragé et conseillé tout au long de ce stage.

Je tiens à montrer tout ma gratitude envers toutes les personnes qui ont pu m'aider, m'encourager, me soutenir, me remotiver pendant ces années de travail.

## 1 Résumé

Pendant ces quatre mois de stage, notre groupe de recherche travaille avec les employés de CMCC ( China Mobile Communications Corporation ) . Le objectif du sujet est : utilise les technique de Fouille de données, étudie les données fournir par le CMCC, et trouve les relation entre les données et les défaut du système 4G. Nous avons fait plusieurs tentatives pour trouver les résultats, et on a utilise différents logiciel, j'ai utilisé le R, et mon collègue utilise Matlab, nous avons utilisé plusieurs algorithmes (Clustering, PCA, Association rules, Ajustement). Mais à la fin, nous avons trouvé que à cause des défaut dans la système d'acquisition, les données ne sont pas correct, et nous ne pouvons pas trouver le résultat comme prévu. Mais les recherches que nous avons fait peut faites-leur savoir comment utilise les technique de fouille de donnée dans la domaine de télécommunication.

## 2 Introduction

Le 3 avril 1973, M. Mation COOPER le directeur général de la division communication de Motorola, à effectuer un appel téléphonique à Joel ENGEL, son rival et néanmoins confrère chez Belle Labs. c'est la premier appel téléphonique en extérieur, L'idée du téléphone portable devient une réalité.

depuis ce jour, le technique développé très rapidement. dans les 20 dernières années, il y a déjà quatre génération des standards pour la téléphonie mobile, non seulement nous pouvons appeler les autres, les nouvelles technologies et les Smart-phones nous permettons aussi envoyer les message, surfer l'Internet, utiliser le service RTSP(Real Time Streaming Protocol), et le service VoIP (Voice over Internet Protocole),etc.. les services de communication téléphonique sont devenus un outil très important dans notre vie.

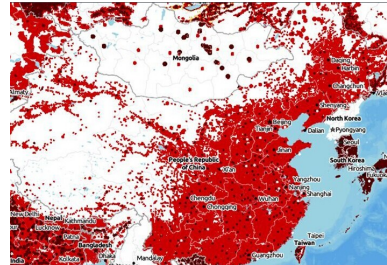
### 2.1 Introduction du CMCC

Fondé en 3 Septembre 1997, après le regroupement de opérateur des télécommunications en 2008, CHINA MOBILE COMMUNICATIONS CORPORATION (CMCC)[1\(a\)](#) est devenu un de trois opérateur des télécommunications en Chine (deux autres sont China Unicom Co., Ltd. et China Telecom). Après plusieurs années de développement, il a construit le plus grand réseau de communications mobiles dans le monde, possède la plus grande base d'utilisateurs dans le monde[1\(b\)](#). En 2013, le CMCC a 767 million utilisateurs,

630,2 billion ¥ de revenu, 121,7 billions ¥ de revenus net, effectif 197,030.



(a) Logo de China Mobile



(b) Réseau télécommunication

FIGURE 1 – CMCC

## 2.2 La crise du CMCC

### 2.2.1 La demande de mise à niveau du réseau

Mais en même temps, le taux de croissance des nouveaux utilisateur décline de 22,5 % (2006) à moins de 5% 2013 [2](#). Et dans la premier 3 mois, l'entreprise une fois considérés comme la plus rentable de Chine, le taux de croissance des revenu net est 0,3%.

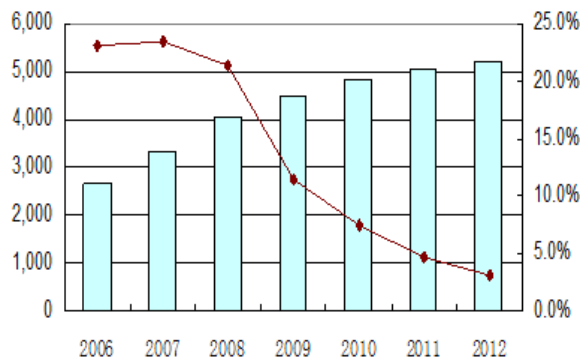
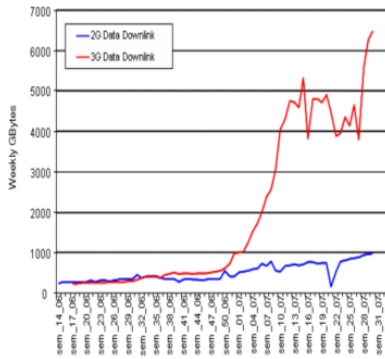


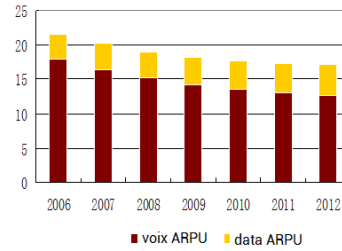
FIGURE 2 – le taux de croissance est décliner

Opérateur des télécommunications Vodafone a fait un étude après il déployé un réseau 3G(the third generation of mobile phone mobile communication technology standards). Comme le réseau 3G permettant des débits (de 2 à 42 Mb/s définis par la dernière génération des réseaux) qui sont bien plus rapides que la génération précédente, par exemple le GSM. Les utilisateur utilisent bien plus souvent le service internet<sup>3(a)</sup>. Comme ils utilisent plus du

service internet, le data ARPU (Average Revenue Per User) augment, mais le voix ARPU décline plus rapide que la montant de data ARPU<sup>3(b)</sup>.



(a) Downlink Data Traffic in 2G/3G Network



(b) étude de Vodafone

FIGURE 3 – Vodafone

Mais l'étude de Orange nous montre que si nous pouvons fournir des nouveaux technologies qui a plus haute débit, les utilisateur utiliseront plus souvent le service data. <sup>4</sup>

Traffic per user per technology used

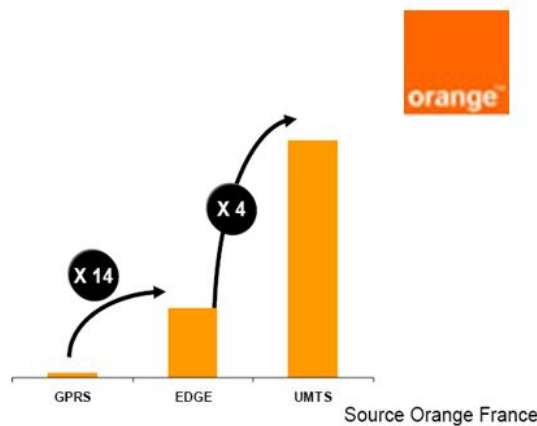


FIGURE 4 – trafic par personne

Des études nous montre nouveaux technologie (comme LTE) peut diminuer le prix de revient, qui peut assurer le profit de l'opérateur. Mais déployer les

nouveaux matériel coût très cher, en 2009, le CMCC dépenser 30 billions ¥ en construit les station pour réseau 3G, et à 2014, le CMCC a construit 1,5 million stations, à la fin de cet année, il y aura 1,8 million stations, parmi ces station, il y aura 500 mille stations TD-LTE. En ajoutant des équipement 4G, il peut être mis à niveau un station de 3G à 4G. Donc déployé le réseau 4G n'est pas trop cher, selon l'expérience précédente (de 2G à 3G), les utilisateurs iront utiliser plus le service internet, qui peut assurer le profit de l'entreprise.

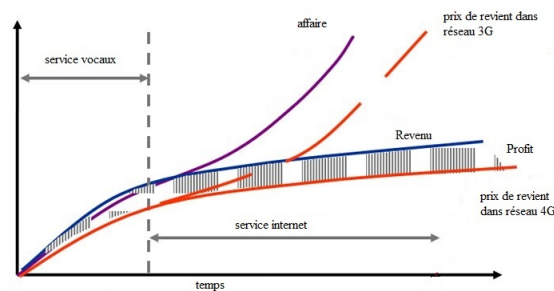


FIGURE 5 – 4G est plus rentable

### 2.2.2 Les changements des moyennes de revenu

Avant la popularité des téléphones intelligents, et avant la popularité du réseau 3G, beaucoup des utilisateurs du CMCC utilisent le services message pour recevoir des informations comme es nouvelles, les météo, etc. Ces services étaient un moyen de revenu très important pour le CMCC. Mais maintenant, avec la popularité du smartphone et l'évolution du réseau, les gens peuvent trouver tout les informations sur l'internet. En la vacance du nouvel an chinois 2013, 31,3 milliards SMS a été envoyé, comparé à cette année, le nombre est 18,2 milliards SMS, il a réduit 42%. Ces informations sont très inquiétantes pour le CMCC.

En conclure, CMCC devez mettre à jour son réseau de télécommunication, mais le mise à niveau du réseau à une réduction de revenu des autres services. Alors CMCC veut apprendre les entreprise comme Tencent et trouver des moyens pour augmenter le revenu.



## 2.3 L'optimisation du réseau

A part de la évolution des technologies. Un grand enjeu pour les opérateurs est : l'optimisation du réseau télécommunication. Le réseau de communication mobile est très dynamique, la répartition de la densité du trafic est inégale, fréquence très limité, etc. La configuration du réseau état toujours sous-optimal, et la perception de l'utilisateur n'est pas très bien. Donc tous les opérateurs doivent toujours reconfigurer/optimiser/maintien les paramètres du réseau.

Les opérateurs peuvent percevoir les données sur Internet, et utilisent ces informations pour trouver les défauts du système, peut aide l'entreprise optimiser le système.

Mais la optimisation du réseau télécommunication est difficile parce-que : Les technologies d'optimisation de réseau concerné : La technologie de commutation, la technologie sans fil, la configuration et commutation de la fréquence, la signalisation système, l'analyse de trafic, etc. c'est un travail difficile, exiger une meilleure aptitude des employés.

Actuellement, l'optimisation du réseau dépend principalement à la expérience du personnel. Mais des fois les expériences ne sont pas correct. Par exemple, Si l'entreprise besoin de savoir le congestionné d'un station, il faut envoyer les employé avec des équipement pendant les périodes de pointe, mais on ne sait pas si les résultats sont correct <sup>6</sup>. En outre, souvent un seul type de donnée ont utilise pour l'analyse et la comparaison pour optimiser les réseau, plutôt que de trouver un solution d'optimisation basées sur toutes les données liées au réseau (telles que les données statistique de trafic, les données d'essai, etc). Et en raison de l'énorme quantité de données, c'est difficile de traite en temps opportun. il est évident que ce méthode est défectueux. Les défauts du système provoque la satisfaction des utilisateurs inférieure, ce qui a conduit à multiplier.

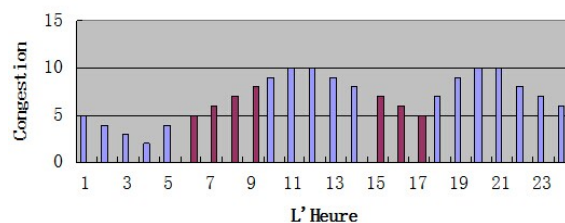


FIGURE 6 – Mesure la congestionné d'un station

Face à des problèmes complexes, les grands entreprises commence utilise

les techniques de Fouille de données. Ce technique peut aide l'entreprise faire les décision plus vite et plus précis.

De ce faire, en Juillet 2013, le CMCC a lancé ce projet avec quatre laboratoires dans trois université, ils sont [Tsinghua University](#), [Shandong University](#) et [University of Electronic Science and Technology of China](#). Le projet inclure trois partiel : Fouille de données, gérés le Cloud plateforme et modélisation de l'information dans le système.

## 2.4 Introduction du laboratoire

De 20 Avril 2014 à 20 Juillet 2014, je fait mon stage chez [laboratoire of Next Generation Network Technology & Application \(NGN\)](#) 7. C'est d'un subordonné de [Research Institute of Network And Human-Machine Speech Communication](#), Département Ingénierie électronique, Tsinghua University. Le laboratoire se trouve dans la ROHM bâtiment.



FIGURE 7 – Logo NGN

Le principaux axes de recherche sont Théorie des réseaux, Architecture de l'Internet, Traitement de l'information Internet, La recherche dans le domaine de la sécurité Internet, Sentiment analyse, Information hiding, etc.

Mon tuteur professionnel est [M. Yongfeng HUANG](#), vice-directeur de la laboratoire NGN. Dans le laboratoire, il y a cinq groupe, chaque groupe a un docteur et son sujet. dans notre groupes, il y a trois personnes, un étudiant de premier année docteur, un étudiant de M1, et une étudiante de Licence troisième année. On utilise R et Rstudio, et Hadoop aussi.

## 2.5 Objectif du projet

Dans cet article, nous avons d'abord présente le réseau communication mobile, ensuite je vais décrire l'état de l'optimisation du réseau et les techniques pour l'optimisé. Enfin je présente notre solution et le conclusion de ce stage.

### 3 Introduction de l'industrie de la télécommunication

#### 3.1 L'évolution des normes de téléphonie mobile

Depuis 1984, il y a déjà plusieurs standards ont été utilisé par les opérateur dans le monde entier. Voici un tableau de différentes standards mobile en Europe et ses paramétrés [1](#).

Génération	Acronyme	Description	Débit
1G	Radiocom 2000	Échanges de type voix uniquement	analogique
2G	GSM	Échanges de type voix uniquement	9,05 kbps
2,5G	GPRS	Échange de données sauf voix	171,2 kbps / 50 kbps / 17,9 kbps
3G	UMTS	Voix + données	144 kbps rurale, 384 kbps urbaine, 1,9 Mbps point fixe / -
3.5G ou 3G+ ou H	HSPA	Évolution de l'UMTS	14,4 Mbps / 3,6 Mbps / -
4G	LTE	Long Term Evolution (Données)	150 Mbps / 40 Mbps / -
4G	LTE-Advanced	Long Term Evolution Advanced (Données+voix)	1 Gbps à l'arrêt, 100 Mbps en mouvement / - / -

TABLE 1 – Les différentes générations de téléphonie mobile en Europe

### 3.1.1 La premier génération

En télécommunication, 1G est la premier génération des standards pour la téléphonie mobile, il s'agit de la première apparition du réseaux de téléphonie mobile, 1G sont des réseaux analogiques, peut échanges de type voix uniquement.

### 3.1.2 La deuxième génération

2G, la technologie de téléphonie sans fil de deuxième génération, la différence entre le réseaux 1G et 2G est : le signaux radio sur les réseaux 1G sont analogiques, et celle de 2G sont numériques.

Systèmes 2G ont été significativement plus efficaces du spectre permettant de bien plus grand taux de pénétration du téléphone mobile, en plus les données vocales numériques peuvent être compressées et multiplexées beaucoup plus efficacement que les codages de la voix analogique grâce à l'utilisation de codecs différents, ce qui permet plus d'appels à transmettre dans la même quantité de bande passante radio. Et 2G présenté premier foi le service de données pour mobile. Les Technologie 2G permettent les divers réseaux de téléphonie mobile de utiliser des services tels que le SMS et MMS. Tous les message de texte envoyés au delà de 2G sont chiffrés numériquement, ce qui permet le transfert de données de telle sorte que seul le destinataire peut recevoir et lire.

Réseaux 2G ont été construits principalement pour le service téléphoniques et de transmission de données lent (défini dans les documents de spécifications IMT-2000).

Réseaux 2,5G, on le qualifie souvent de le General packet Radio Service ou GPRS, est une norme pour la téléphonie mobile dérivée du GSM et complémentaire de celui-ci, permettant un débit de données plus élevé. Le 2,5 indique que c'est une technologie à mi-chemin entre le GSM (deuxième génération) et l'UMTS (troisième génération). Le GPRS est une extension du protocole GSM : il ajoute par rapport à ce dernier la transmission par paquets. Cette méthode est plus adaptée à la transmission des données. En effet, les ressources ne sont allouées que lorsque des données sont échangées, contrairement au mode « circuit » en GSM où un circuit est établi – et les ressources associées – pour toute la durée de la communication. Le GPRS a ensuite évolué au début des années 2000 vers la norme EDGE également optimisée pour transférer des données et qui utilise les mêmes antennes et les mêmes fréquences radio.

### 3.1.3 La troisième génération

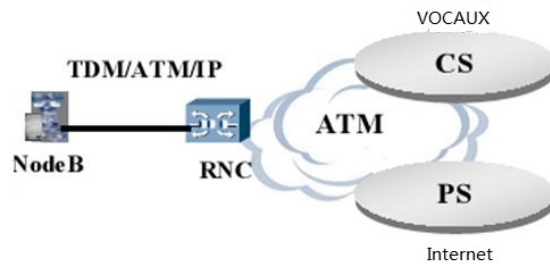
La troisième génération (3G) des normes de téléphonie mobile. Elle est représentée principalement par W-CDMA, CDMA2000, TD-SCDMA et WiMAX. Elle permettant des débits de 2 à 42 Mb/s qui sont bien plus rapides qu'avec la génération précédente. Grâce à l'utilisation des règles de classement de l'utilisateur, et les bandes de fréquences supérieures rendant la capacité du réseau augmenter.

Dans les différents standards 3G et ses prédécesseurs, ils utilisent le domaine CS (Circuit Switch) pour le service vocal, et le domaine PS (Packet Switch) s'occupe du service de données [8\(a\)](#).

### 3.1.4 La quatrième génération

La quatrième génération des standards pour la téléphonie mobile, succédant à la 2G et la 3G, en théorie, elle permet de transmettre de données à des débits supérieurs à 100 Mb/s.

Une des particularités de la 4G est son EPC (Evolved Packet Core) basé sur IP, et il n'y a plus de mode commuté (le 'Circuit Switched Domain' qui s'occupe du service vocal dans les standards précédents), ce qui signifie que le service vocal transmis sur l'internet [8\(b\)](#).



(a) Réseau 3G et ses prédécesseur



(b) Réseau 4G

FIGURE 8 – Structure des réseaux

Les avantages du réseau 4G sont : plus haut débit, mieux utilisation de la bandes de fréquence, moins de délai (délai dans le panneau de l'utilisateur est inférieur que 5 ms, délai dans le panneau de commande est inférieur que 100 ms ), plus simple structure du réseau, moins de consommation d'énergie Terminal.

### 3.2 Le réseau LTE

Le LTE (Long Term Evolution) est l'évolution la plus récent des normes de CDMA 2000, TD-SCDMA, GSM. La norme LTE. La technologie LTE été considérée comme une norme de troisième génération '3.9G', et la 'vraie 4G', appelée LTE-Advanced été reconnu par l'UIT comme une technologie 4G en 2010. LTE a deux branche : LTE-FDD (Frequency-Division Duplex Long Term Evolution)et LTE-TDD, (Time Division Duplex Long Term Evolution)les deux standards sont similaire, la différence entre les deux standard est moins de 15% <sup>9</sup>. En 2011-2012, les réseaux LTE-TDD sont commercialisés sous l'appellation 4G par le CMCC un Chine.



FIGURE 9 – l'évolution des standard

### 3.2.1 La structure du réseau LTE

Le réseau 4G contient 3 partie : UE( User Equipment);, eNodeB (les stations de base), EPC (Evolved Packet Core). EPC contient MME, S-GW, P-GW et HSS 10 2.

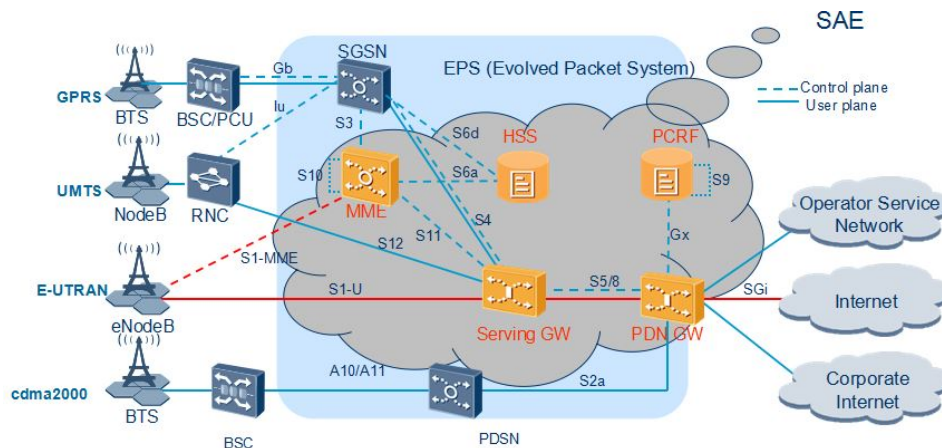


FIGURE 10 – la structure du réseau

Part	Fonction
MME	L'authentification des utilisateurs et la gestion des clés, Cryptage de la couche NAS, Gestion de la liste TA, Sélection P-GW ou S-GW
Service Gateway	Compression d'en-tête IP, Routage de paquets et la transmission, La commutation entre eNB, Facturation des utilisateurs porteur
PDN Gateway	L'allocation des adresses IP de UE, l'accès aux fonction de gestion de réseau externes, Facturation en service
HSS(Home Subscriber Service)	Stockée données de l'utilisateur associées au service
PCRF	Roaming

TABLE 2 – la fonction du chaque partie

Entre deux E-UTRAN, il y a l'interface X2, l'interface S-11 se trouve entre S-GW et MME, E-UTRAN et S-GW échange les données par l'interface S1-U et il échange les donnée par l'interface S1-AP avec MME, MME et HSS utilise l'interface S6A, et l'interface S5/8 entre S-GW et P-GW, Gx entre PCRF et P-GW. En mettant des capteur en les interfaces, les opérateurs et les fournisseurs d'équipement peuvent collecter les données de signalisation, et utilisent ces informations pour trouver les défauts du système.

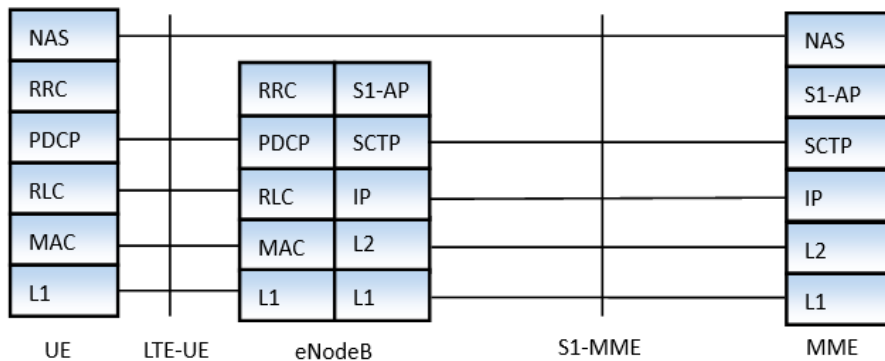


FIGURE 11 – Contrôle plan



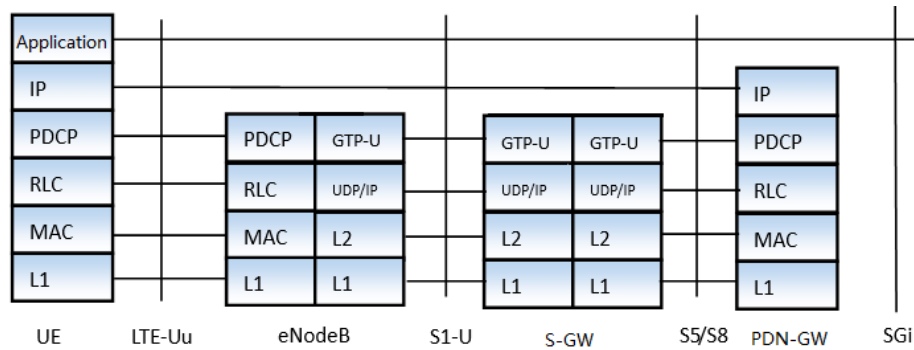


FIGURE 12 – User plan

## 4 Les solution existant

L'optimisation du service téléphonie est très important. L'opérateur a construit un immense réseau télécommunication, mais a cause de la mauvaise configuration du système, les utilisateur ne sont pas satisfaits avec les services, les investissement n'a pas été remis. Donc les entreprises comme IBM, Huawei, et l'autre fournisseur du équipement essaient de trouver la meilleure solution.

Maintenant, il y a beaucoup des gens travaillent sur ce domaine, nous avons trouvé beaucoup de papier sur l'optimisation du réseau télécommunication, mais les articles sont basé sur réseau 3G ou 2G.

Il y a trois techniques qui sont beaucoup utilisé :

1. la Technique KQI ;
2. la Technique QoE ;
3. la Technique qui étudie les comportements de l'utilisateur.

la Technique KQI :

La technique utilise le plus souvent s'appelle 'KQI' ( Key Quality Indicator) [13](#), cette méthode a été beaucoup utilisé. Et cette technique peut généralement divisé en deux étapes. d'abord, nous devons calculer le score de KPI, pour calcule le KPI en premier, il faut analyser le processus d'un service et choisir les indicateur de performances. Ensuite, nous pouvons calculer le score d'un processus en utilisant un équation linéaire, le poids de chaque attributs change selon le service, par exemple, pour le service SMS, le délai porte peu de l'importance, mais le délai du service est important pour le service HTTP. à la fin, nous pouvons calculer le KQI à avec les KPI [\[2\]](#).

Mais les poids sont défini par les experts, et les valeurs peut-être fausse ou pas précis. Et par fois le score est bonne mais l'expérience de l'utilisateur n'est pas bon.

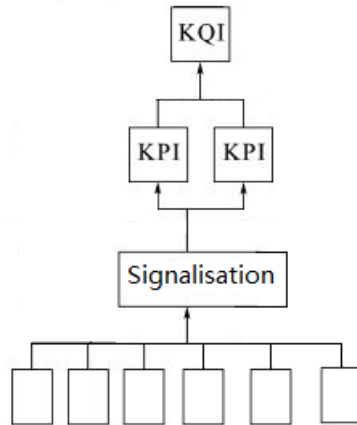


FIGURE 13 – KQI

#### la Technique QoE :

KPI est un des indicateurs de qualité axés sur les performances du réseau, mais il ne reflète pas directement l'expérience de la qualité de service de l'utilisateur, parce que les expériences de l'utilisateur sont difficile à mesure. Donc la technique QoE a été inventé. QoE défini la performance, de la qualité de service et l'expérience de l'utilisateur de l'ensemble du réseau à partir de l'utilisateur.

Les utilisateurs ont nombreuses exigences pour les services téléphonie, ils peuvent être résumées comme deux aspects : la fiabilité et le confort. La fiabilité fait référence à l'activité de l'accessibilité, la disponibilité et la durabilité. Le confort est une qualité de service, est un indice de la perception directe de l'utilisateur, qui dépend à l'expérience de l'utilisateur[3]. Les relations entre QoE et QoS KPI sont : la fiabilité du service 3, le confort du service4.

TABLE 3 – Fiabilité du service

KQI	QoE
Accessibilité	Taux de succès
Disponibilité	Temps d'accès aux services
Durabilité	La durée de l'accès des services

TABLE 4 – Confort du service

KQI	QoE
	Taux de perte de paquets de couche d'application
	Le débit moyen
La qualité du service de transmission	Stabilité de la transmission
	Le bout en bout délai moyen
	Gigue
Le persistant de la connexion de service	La vitesse et la difficulté du service d'assistance

Maintenant, la technique de QoE a été beaucoup utilisé pour le service vocaux. Et grâce à la complexité du service de données, il n'y a pas un standard de QoE pour le service de données.

En utilisant la technique KQI et QoE, nous peuvent mesurer la qualité du service, les résultats peuvent aider les opérateurs trouver les services de mauvaise qualité, les opérateur peuvent améliorer les services selon le résultat, finalement améliorer la notation de l'utilisateur.

Le résultat de KQI dépend seulement aux performances du réseau, donc nous avons besoin les informations des performances de réseau. Et la technologie QoE besoin Le résultat de KQI et les feed-back de l'utilisateur, le feed-back peut obtenir par l'enquête ou les plaintes des utilisateurs, et par les mesures directs.

la Technique qui étudie les comportements de l'utilisateur

Aussi il y a un groupe qui utilise le comportement de l'utilisateur pour défini la qualité du service[4]. Le groupe utilise cette méthode dans le service vocaux, il cherche le situation comme l'utilisateur accroche et ré-appel le même personne. À la fin, cette méthode aide l'opérateur corriger le paramètre d'erreur.

Selon l'article le cette méthode peut aider l'opérateur trouver les défaut du système, mais il aa nombreuses restrictions, par exemple, nous ne pouvons pas utiliser cette technologie dans le service de SMS, etc.

## 5 Le présentation de notre solution

La méthode qui utilise le comportement de l'utilisateur est intéressant, mais nous trouvons qu'elle peut utiliser seulement dans le service vocaux, nous n'avons pas trouvé les règles similaire dans l'autre service. D'ailleurs, le réseau LTE ne support pas le service vocaux, donc il n'existe pas optimisation du service vocaux dans réseau LTE et nous n'avons pas de données. Donc nous ne pouvons pas utiliser cette méthode.

La technique QoE et KQI sont beaucoup utilisé, mais d'abord, pour la méthode QoE, nous avons besoin les réponses des utilisateurs, mais nous n'avons pas assez de temps, et des raisons financières, le CMCC ne peut pas nous fournir ces données. Et le équation qu'on utilise pour calcule KPI ne sont pas convaincante, voici un exemple d'un équation pour calcule la disponibilité du réseau pour le service SMS dans réseau 3G14.

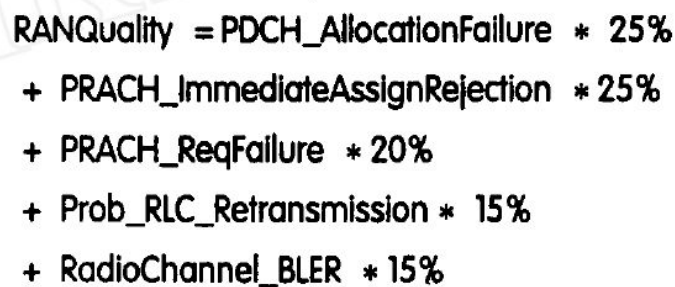

$$\begin{aligned} \text{RANQuality} &= \text{PDCH\_AllocationFailure} * 25\% \\ &+ \text{PRACH\_ImmediateAssignRejection} * 25\% \\ &+ \text{PRACH\_ReqFailure} * 20\% \\ &+ \text{Prob\_RLC\_Retransmission} * 15\% \\ &+ \text{RadioChannel\_BLER} * 15\% \end{aligned}$$

FIGURE 14 – Un exemple d'un équation pour calculer la disponibilité

Le poids du chaque attributs sont définir par les experts, Mais l'utilisateur n'est pas satisfaits du service. Nous croyons que l'erreur a été causée par l'inexacte équation, et nous pensons que les algorithmes de Classification peuvent aider à améliorer le résultat, mais très vite nous avons trouver que

le CMCC ne peut pas nous fournir ce type de données. Sans la connaissance a priori, nous ne pouvons pas utiliser ces algorithmes. Nous avons aussi pensé à utiliser l'externalisation ouverte (crowdsourcing), à notre avis si le CMCC peut lancer un projet de externalisation ouverte, si le CMCC peut encourager ses utilisateurs donnent les notes aux services pour obtenir des crédits, nous pouvons obtenir la connaissance a priori, et à l'aide de ces données, nous pouvons trouver une équation peut-être mieux que les équations écrites par les experts. Mais bien sûr, le CMCC n'a pas accepté cette idée, parce que cette méthode peut coûter cher, et peut-être il n'y a pas de revenu direct. Et l'entreprise ne fait pas de l'investissement sans retour. Donc nous n'avons pas de connaissance a priori.

Finalement, nous avons décidé d'utiliser la technique de l'apprentissage non supervisé, il contient la notion de Réseau de neurones et l'algorithme de clustering etc. Nous avons choisi l'algorithme de clustering. Et nous avons utilisé la technique de l'arbre couvrant de poids minimal en anglais Minimum spanning tree (MST) et la méthode Règles d'association et PCA.

## 5.1 Clustering

L'algorithme de clustering est une des méthodes de classification non supervisée. Il est beaucoup utilisé quand la donnée n'a pas de connaissance a priori.

C'est une méthode statistique d'analyse des données. Elle divise un ensemble de données en différents groupes, les données de chaque groupe ont mathématiquement plus de proximité que les données de l'autre groupe, et nous supposons que les données dans la même partition ont des caractéristiques similaires.

Il existe de multiples méthodes de regroupement des données, parmi lesquelles :

- Classification basée sur la densité ;
- Classification hiérarchique ;
- Classification par partitionnement ;
- Classification par grille ;
- Classification basée sur des modèles.

Les étapes de cet algorithme sont :

- Choisir  $k$  points qui représentent la position moyenne des  $k$  partitions initiales (au hasard) ;
- Répéter les étapes suivantes jusqu'à convergence :

1. assigner chaque observation à la partition la plus proche
2. mettre à jour la moyenne de chaque cluster

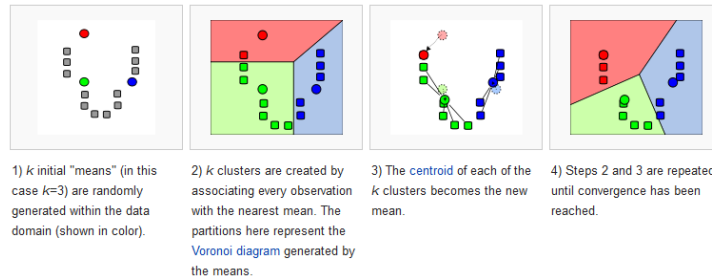


FIGURE 15 – étapes de l'algorithme K-Means

Nous avons utiliser l'algorithme K-Means et CLARA (Clustering Large Applications ). Ils sont deux techniques de Classification par partitionnement.

Le but de cet algorithme est de diviser des données en  $K$  partitions (clusters) dans lesquelles les données appartient à la partition avec la moyenne la plus proche.

### 5.1.1 La distance

Quand on regroupe un ensemble de données, on calcule la distance entre chaque échantillon, la distance entre les échantillons dans un même groupe sont plus petit que les échantillons dans l'autre groupe. on suppose que les plus semblables les deux échantillons, plus la distance et petit, plus la similitude est grand.

Il y a plusieurs moyen de calculer la similarité, On utilise la technique qui basé sur la distance, les techniques utilisé le plus souvent sont : *distance euclidienne*, *distance de Hamming*, *distance Manhattan*, etc. Aussi il existe des méthodes qui mesuré la similarité, il inclut : *coefficient de Jaccard*, *cosinus similarité*, etc . Nous décidons de utiliser la distance euclidienne pour calculer la similarité.

$$d(x, y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2}$$

$n$  est la dimension,  $x_k$  et  $y_k$  sont le  $k$ -ème attributs de échantillon  $x$  et  $y$ .

### 5.1.2 La validité de clustering

C'est méthode a été utilisés dans beaucoup différents domaines. Mais la qualité du résultat dépend au nombre de clusters ou la valeur de K, différent paramètre peuvent mener à des résultat très différent.

En 1974, Mme BEZDEK a proposé cette question de évaluation, et donne la première fonction pour évaluer le résultat du clustering : **Partition Coefficient**. Par la suite de nombreux chercheurs ont proposé une variété de fonctions pour évaluer le résultat, et évalue les qualité et le champ d'application de ses fonctions.

Par mesure l'efficacité entre les partitions et inter-partition, on peut évaluer le clustering. Le résultat de clustering idéal devrait être avoir la distance minimale dans le partitions, et la distance maximale entre les partitions. C'est à dire avec le plus petit grand de la cohésion dans la partition, et le plus élevé degré de la séparation entre les partitions. La cohésion mesure le degré de proximité dans la groupe, et la séparation mesure le degré de dissimilarité entre les groupe. La cohésion et la séparation peuvent calculer par les équations suivante.

$$cohson(C_i) = \sum_{x \in 1} proximity(x, c_i)$$

$$sparation(C_i, C_j) = proximity(c_i, c_j)$$

Dans les formules, le  $c$  est le centre de gravité d'un partition,  $proximity(a, b)$  est le mesure de proximité entre la partition a et b.

Le degré de la cohésion de clusters et de la séparation est souvent utilisé comme une mesure principales pour évalue le résultat de clustering. Un bon regroupement devrait être à la fois un petit degré de la cohésion et un grand degré de la séparation.

Le degré de la cohésion et de la séparation de la partition ne sont pas indépendants, la somme des deux est une constante, on suppose que quand on a le minimal degré de la cohésion, on peut avoir le maximal degré de la séparation. Évidement, nous devons utiliser à la fois le degré de la cohésion et de la séparation pour mesuré la qualité du regroupement.

### 5.1.3 La silhouette Coefficient

Kaufman a proposé la silhouette coefficient en 2010, cette méthode utilise à la fois les deux degrés.

1. la silhouette coefficient d'un échantillon :

Pour un échantillon  $d_i$ , en supposant qu'il est dans groupe A, la silhouette coefficient peut calculer par ce formulaire :

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

$a_i$  est la dissemblance moyenne de la donnée  $i$  avec toutes les autres données dans le même cluster (plus la valeur est petit, meilleure est le regroupement).  $b_i$  est la différence moyenne de  $i$  à tout autre groupe  $i$  n'est pas un membre.

Qui peut être écrite comme :

$$s_i = \begin{cases} 1 - \frac{a_i}{b_i}, & \text{if } a_i < b_i \\ 0, & \text{if } a_i = b_i \\ \frac{b_i}{a_i} - 1, & \text{if } a_i > b_i \end{cases}$$

D'après la définition ci-dessus, il est clair que la valeur de  $s_i$  est entre 1 et  $-1$

$$-1 < s_i < 1$$

Pour  $s_i$  d'être proche de 1, nous demandons le  $a_i$  plus petit que  $b_i$ . Comme  $a_i$  mesure la dissemblance de  $i$  et son propre groupe, une petit valeur signifie qu'il est bien adapté. En outre, un grand  $b_i$  implique que  $i$  est mal adapté à son groupe voisin. Ainsi, un  $s_i$  près de 1 signifie que la donnée est concentrée de manière appropriée. Si la valeur de  $s_i$  est proche de  $-1$ , selon la même logique, nous voyons que le donnée  $i$  serait plus approprié si elle a été regroupée dans son groupe voisin.

2. la silhouette coefficient moyenne :

Pour le résultat d'un regroupement, le silhouette coefficient égale à :

$$s_k = \frac{1}{n} \sum_{i=1}^n s_i$$

$n$  est le nombre de données,  $k$  est le nombre de partition,  $s_k$  est la moyenne du silhouette coefficient. Nous pouvons utiliser  $s_k$  pour évalue



*la qualité du clustering.*

Le  $s_i$  est les moyens sur l'ensemble des données d'un cluster, il est une mesure du degré de cohésion de toutes les données du cluster. Ainsi, les moyens de  $s_i$  de l'ensemble des données ( $s_k$ ) est une mesure de la façon appropriée les données ont été regroupées. S'il y a trop peu de clusters, ainsi que peut se produire lorsque un mauvais choix de  $k$  est utilisé dans l'algorithme K-Means, la silhouette Coefficient de certains groupes est beaucoup plus petit que les autres. Donc la plot de silhouette Coefficient et la valeur moyenne de silhouette peuvent être utilisés pour déterminer le nombre de cluster (la valeur de  $k$ ) optimal pour l'ensemble de données.

Après trouver les paramètres optimaux, nous pouvons utiliser l'algorithme clustering, et après étudié et comparé les caractéristiques de chaque partition, on peut définir quelle partition représente une mauvaise qualité de service. Le résultat peut nous aider à classer le service.

## 5.2 Règles d'association

La règle d'association est une méthode populaire, elle est étudiée d'une manière approfondie dont le but est de découvrir des relations ayant un intérêt pour le statisticien entre les variables. En se basant sur le concept de relations fortes, Rakesh Agrawal et son équipe présente des règles d'association dont le but est de découvrir des similitudes entre des produits dans des données saisies sur une grande échelle dans les systèmes informatiques des points de ventes des chaînes de supermarchés. Par exemple, une règle découverte que si un homme achète les serviettes de bébé, il est susceptible d'acheter les bières. Une telle information peut être utile quand on veut prendre des décisions marketing.

Les règles d'association sont employées aujourd'hui dans plusieurs domaines, incluant : la fouille du web, la détection d'intrusion et la bio-informatique. Dans ces domaines, ils utilisent les données booléennes pour trouver les règles utiles. Mais dans le domaine télécommunication, les données de signalisation sont les données numériques. Donc nous ne pouvons pas directement utiliser les règles d'association. Par contre nous pouvons convertir les données de type numérique en données de caractère en divisant les données dans plusieurs partitions [6]. Par exemple, nous avons des données numériques entre 0 et 100, et on divise les données en 10 partitions, donc les chiffres entre 0 et 10 peuvent représenter par  $0 < x < 10$ . En utilisant

cette technique nous pouvons utiliser la règle d'association pour trouver les relations dans les données.

Il y a plusieurs méthodes pour diviser les attributs numériques : par catégorisation, l'analyse typologique, par analyse de l'histogramme, l'analyse basé sur l'entropie de discret, partition naturel et ainsi de suite.

Nous décidons de utiliser le K-Means d'abord. Après analyse le résultat, nous pouvons utiliser la algorithme règles d'associations à l'aide du résultat de K-Means.

### 5.3 K-Means et l'arbre couvrant de poids minimal

D'après les techniques précédant, nous avons utilisé l'arbre couvrant de poids minimal avec K-Means.

En théorie des graphes, étant donné un graphe non orienté connexe dont les arêtes sont pondérées, un arbre couvrant de poids minimal de ce graphe est un arbre couvrant (sous-ensemble qui est un arbre et qui connecte tous les sommets ensemble) dont la somme des poids des arêtes est minimale<sup>16</sup>. L'arbre couvrant de poids minimal est aussi connu sous certains autres noms, tel qu'arbre couvrant minimum ou encore arbre sous-tendant minimum. L'algorithme de Prim et l'algorithme de Kruskal sont deux méthodes classique, qui sont tous les deux l'algorithme glouton.

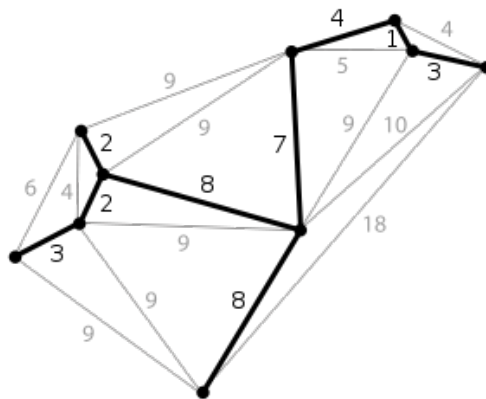


FIGURE 16 – MST

En coupant  $C - 1$  arêtes le plus grand, nous pouvons grouper le données en  $C$  parties.

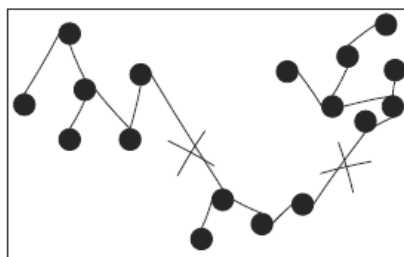


FIGURE 17 – MST clustering

Dans notre projet, nous avons choisi l'algorithme Prim et K-Means pour faire le clustering.

La complexité de l'algorithme Prim est :  $O((V + E) \lg(V)) = O(E \lg(V))$ ,  $V$  est l'ensemble de vertex,  $E$  est l'ensemble de arêtes.

la complexité de l'algorithme K-Means est :  $O(nkt)$ ,  $n$  est la quantité de données,  $k$  est le nombre de partitions,  $t$  est le nombre d'itérations.

### 5.3.1 L'algorithme KmMST

Dans l'article [7], nous avons trouvé un algorithme qui utilise K-Means et la MST.

Parce que K-Means peut trouver des partitions sur forme sphéricité, et la valeur de  $K$  n'est pas grand donc le résultat est influencé par les données de bruits, donc le résultat de K-Means est parfois peu satisfaisant.

D'après l'article, nous pouvons choisir une grande valeur pour  $K$ , puis utiliser la technique MST pour combiner la partition en  $C$  groupes en coupant  $C-1$  arête. La performance est mieux que celle de K-Means [18].

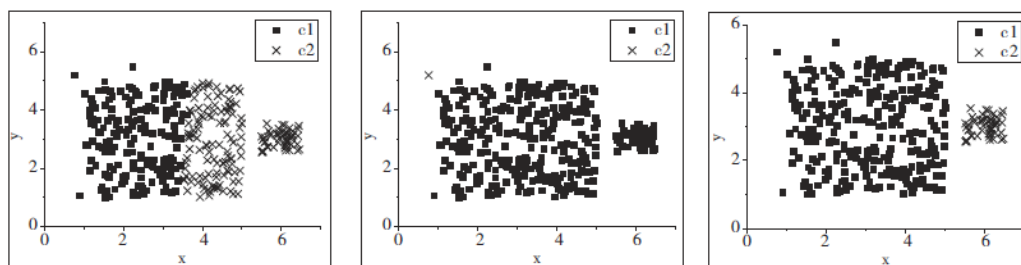


FIGURE 18 – La performance de K-Means, MST et KmMST

L'algorithme KmMST peut décrire comme suit : Input : le données de quantité  $n$ , coefficient  $r$  et  $C$ . output : les partitions. l'étape :

1. calcule la valeur de  $K$ ,  $k = \lfloor n^r \rfloor$  ( $r \in [0, 1]$ ) par défaut la valeur de  $r$  égale à 0,5 ;
2. utilise l'algorithme K-Means, trouve  $K$  partitions ;
3. calcule la distance entre chaque centre de la partition ;
4. utilise l'algorithme Prim pour crée l'arbre couvrant de poids minimal ;
5. coupe le  $C - 1$  plus bords qui ont le plus grand distance, le  $C$  sous-graphe est le résultat de clustering.

## 6 La mise en œuvre

### 6.1 Le logiciel utilisés

L'objectif de notre projet est de trouver une méthode qui peut s'implanter dans les serveurs du CMCC, et aider le CMCC améliorer la qualité du réseau. D'après le directeur du R&D département du CMCC, au total, il a 20Pb de donnée stock dans ses base de données, donc le logiciel doit être capable d'exécuter grand quantité de données. En plus, au lieu de utilisé directement les logiciel comme SQL (qui n'est pas très efficient si on a beaucoup de donnée stock dans différent serveur), le CMCC utilise 'Hadoop' pour stock et géré ses données.

Donc le logiciel que nous utilisons doit être capable d'exécuter grand quantité de donnée et peut travailler avec Hadoop.

Finalement nous décidons de utiliser le langage R. L'avantage de R sont :

- R est un langage et un environnement pour le calcule statique et les graphiques ;
- R offre une grande variété de statistiques (modélisation linéaire et non linéaire, classification, clustering, etc,) et des graphiques techniques, et il est très extensible ;
- R est facile à utiliser ;
- R est un logiciel libre, il compile et fonctionne sur une grande variété de plates-formes UNIX et les systèmes similaires (y compris FreeBSD et Linux), Windows et Mac OS ;
- en utilisant les packages fournir par 'Revolution Analytics', nous pouvons utiliser Hadoop en R.

Nous utilisons Rstudio comme notre environnement de programmation. C'est une interface utilisateur puissante et productive pour R.

## 6.2 Introduction des données

Après quelque semaines de négociation avec les employés de différents départements de le CMCC, ils nous ont fourni deux versions de données, et ses spécifications du format[1]. Nous avons trouvé que le CMCC n'a pas de accès direct aux données, et le fournisseur d'équipement a modifié le spécification fournir par le CMCC, et il y a des erreurs dans les données fourni par les fournisseur d'équipement.

Ils nous ont envoyé 11 dossiers, chaque dossier correspond à un service. les services sont 'rtsp', 'dns','mail', 'ftp', 'http-wap', 'mms', 'p2p', 'realtimecom', 'VoIP' et les données de signalisation entre E-UTRAN et MME 'S1AP-NAS'.

Et nous avons trouvé que pour les services comme 'VoIP' et 'RTSP', ils sont très peu de données 5. Donc nous avons décidé de utiliser le donnée du service 'HTTP'.

L'interface	Nombre de ligne
S1-AP	240
RTSP	35
DNS	272562
Maill	44
FTP	71
HTTP-WAP	50854
MMS	193
P2P	515
Realtimecom	2082
S1U	89759
VoIP	28

TABLE 5 – les dossiers de données

Le dossier du service HTTP a 18,4Mbit , il y a 50854 lignes, tous les données sont collectées par les capteurs placer entre les Service-Gateway et les eNodeB. le capteur enregistre un ligne de donnée quand un processus est fini. chaque ligne a 76 attributs19.

data.http 50894 obs. of 76 variables

FIGURE 19 – les données du service HTTP

Il contient des informations de UE (IMEI, IMSI, etc), les trafic de la liaison montante et la liaison descendante et le temps, la adresse IP de UE, eNodeB et S-GW, le port de UE, eNodeB et S-GW, le délai du service, le site web, cookie, et aussi le début temps et le temps d'arrêter. les données sont collectent dans 20.92 minutes 20.

```
StartTimes <-as.POSIXlt(as.numeric(substr(data_HTTP$StartT,1,10)), "UTC", origin="1970-01-01")
EndTimes <-as.POSIXlt(as.numeric(substr(data_HTTP$StopT,1,10)), "UTC", origin="1970-01-01")
RecodeTime <-data.frame(StartTimes,EndTimes)
attach(RecodeTime)
```

```
cat("Data recoded in:",max(EndTimes)-min(StartTimes)," minits")
```

```
## Data recoded in: 20.92 minits
```

FIGURE 20 – Les données sont collectent dans 20.92 minutes

### 6.3 Prétraitement de données

En analysant des données, nous avons trouvé des erreurs de données, et le fournisseur nous a confirmé que ces sont les défaut de leur système 4G. Pour les attributs 'IMSI', 'IMEI', 'MSISDN', 90 % de lignes sont vides, ce qui ne sont pas vides, les contenus sont illisible, et peuvent provoquant des erreurs de lecture 21. Et nous avons trouvé que dans les contenus de certains lignes sont bizarre.

IMSI	IMEI	MSISDN	M.TMSI	IpType
			o  1	175453935
w hw2		hQp p	1	176899993
			H  1	176918718

FIGURE 21 – erreur du codage BCD

Dans ce processus, le 'Down Link Online Time' égale à 0 ms , mais il a téléchargé 746 bits, c'est clairement un erreur du données.

BigType	SubType	L4	ServerIP	ServerPort	UpTraffic	DownTraffic	UpTime	DownTime
15	5017	0	3719544451	80	595	746	500	0

FIGURE 22 – Erreur de la donnée

Nous avons décidé de ne utiliser les données avec ce type de erreur, à la fin, en supprimant ses données, il nous reste 37865 lignes (50894 lignes en origine, 13029 lignes ont été supprimé) [23](#).

data_DisHttp	50894 obs. of 76 variables
data_HTTP	37865 obs. of 76 variables

FIGURE 23 – Prétraitement des données

Entre ces 76 attributs, Une grande partie de ces informations sont inutile, et pour certain attributs les contenus égal tous à 0. Finalement nous avons trouvé 11 attributs. ils sont :

Signalisation	Signalisation	KPI
trafic en liaison montante	le temps en ligne	vitesse
trafic en liaison descendante	le temps en ligne	vitesse
	Http First Response Time	délai
	Http Last Packet Time	délai
	Http Last Ack Time	délai
Packet Num en liaison montante	retransmission de paquets Num en liaison montante	taux de retransmission
Packet Num en liaison descendante	retransmission de paquets Num en liaison descendante	taux de retransmission

Mais nous avons trouvé que dans certains lignes le taux de retransmission sont trop grands ( plus grand que 100%). par exemple, dans un processus, il a téléchargé 17 paquets IP, et il a 7448 paquets sont désordre, et ré-téléchargé 6384 paquets 24. Les données ne sont pas correct, donc nous ne pouvons pas utiliser ses donnée pour calculer le taux de retransmission.

UpPac	DownPac	UpDisPac	DownDisPac	UpRePac	DownRePac
6	5	0	0	0	0
10	19	0	0	0	0
5	4	380	0	0	0
8	17	0	7448	40	6384
3	6	0	2128	0	0

FIGURE 24 – Défaut de la système



Finalement nous avons décidé de utiliser ces 5 attributs (la vitesse et le délai) pour mesurer la qualité du service.

## 6.4 Les caractéristique du donnée

```
> min(upAvBand)      > max(upAvBand)      > sd(upAvBand)
[1] 0.06616667        [1] 170.16          [1] 1.631399
> min(downAvBand)    > max(downAvBand)    > sd(downAvBand)
[1] 0                  [1] 1316             [1] 24.36523
> min(firstRespondTime) > max(firstRespondTime) > sd(firstRespondTime)
[1] 0                  [1] 150695           [1] 4475.789
> min(lastPacketTime) > max(lastPacketTime) > sd(lastPacketTime)
[1] 0                  [1] 92081            [1] 2673.268
> min(lastAckTime)    > max(lastAckTime)    > sd(lastAckTime)
[1] 0                  [1] 1220150          [1] 8989.971
```

FIGURE 25 – La valeur maximal, valeur minimal et l'écart type

```
> mean(upAvBand)
[1] 0.85507
> mean(downAvBand)
[1] 4.735396
> mean(firstRespondTime)
[1] 594.1533
> mean(lastPacketTime)
[1] 392.4637
> mean(lastAckTime)
[1] 1231.584
```

FIGURE 26 – La valeur moyenne

Nous avons trouvé que l'écart type pour les trois attributs de délais sont très grand. Les valeurs varient de  $0ms$  à  $1220150ms$ . Par contre la changement de la vitesse de la liaison montante et descendante n'est pas très grand.

```
> data_PreAR[,3]
[1] 6102 43 0 0 0 249 424 0 0 52 0 10 67 0 0 0 0 30125 57
[20] 23 0 0 0 0 508 0 0 0 0 0 0 0 1249 0 0 0 0
[39] 0 195 0 0 33 0 0 0 0 0 0 0 0 1104 0 0 0 0
[58] 37 0 29 0 66 0 0 0 0 0 0 39 0 0 0 0 0 33
[77] 0 0 236 0 0 0 0 101 0 0 742 0 0 138 0 9 1509 34 669
[96] 0 0 0 0 0 43 0 14 0 35 0 0 42 0 0 0 0 54
[115] 0 0 144 229 69 0 0 0 10 0 0 0 0 100 42 516 0 595
[134] 0 0 0 0 0 360 0 0 0 0 791 0 0 0 0 0 182 0 1271
[153] 0 0 0 0 49572 0 6159 0 25 0 0 0 0 461 1233 0 1431 0
[172] 242 0 0 36 0 76049 31 0 0 24 0 0 0 1347 128 0 0 0
[191] 0 0 0 39 59 0 120 102 1109 0 0 663 0 0 0 26 0 1083
[210] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[229] 0 0 116 0 0 0 0 0 37 0 0 0 1797 0 170 0 0 0
[248] 0 26 0 360 0 0 0 10 151 0 0 207 0 0 0 0 30249 0
```

FIGURE 27 – Le temps de réponse du serveur

En analysant les données, nous avons trouvé que les données ne sont pas normales, et l'ingénieur du fournisseur de l'équipement nous ont confirmé que le système de collecte d'informations de signalisation a des défauts.

## 6.5 Le K-Means et Le Règle d'association

### 6.5.1 La valeur optimal de $k$

Pour trouver le  $k$  optimal pour nos données, nous avons utilisé la technique 'silhouette Coefficient' et 'somme de carrés d'erreur'.

Nous avons mesuré le résultat de ces deux techniques quand la valeur de  $k$  augmente de 1 à 15, étant donné que les caractéristiques de l'algorithme, pour chaque valeur de  $k$ , nous avons répété 50 fois et calculé la valeur moyenne pour éliminer les erreurs, ensuite, nous étudions le résultat pour trouver la valeur optimale.

#### La somme d'erreur carrés

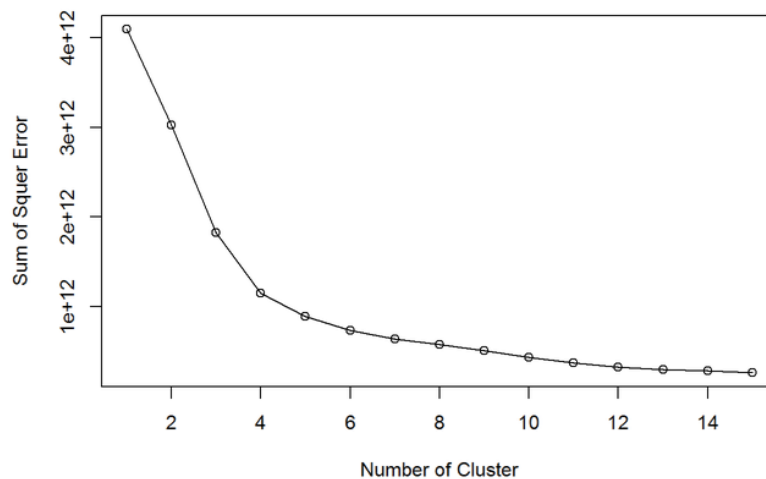


FIGURE 28 – somme d'erreur carrés

Nous avons trouvé que la somme d'erreur carrés diminue quand  $k$  augmente, mais il est difficile de choisir la valeur optimale de  $k$  avec cette image.

### La silhouette Coefficient

Nous utilisons la même technique pour calculer et visualiser la valeur de la silhouette Coefficient.

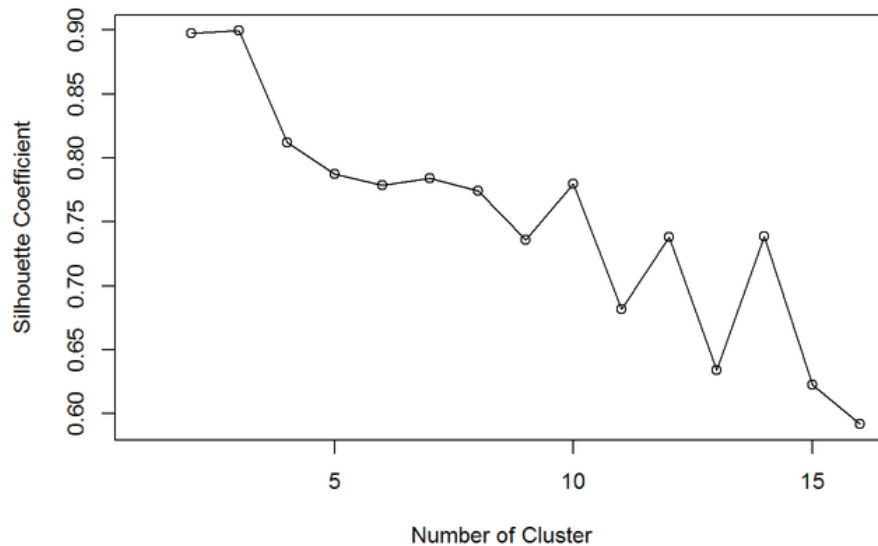


FIGURE 29 – la silhouette Coefficient quand  $k$  varie de 2 à 15

L'image de la silhouette Coefficient nous montrons que comme quand  $k = 3$  la valeur de la silhouette Coefficient est plus grand. Donc nous soupçonnons que quand  $k = 3$  nous pouvons trouver le mieux partition.

#### 6.5.2 Le clustering

Après classifie le données en 3 groupes utilisant l'algorithme CLARA. Nous avons trouvé que la majorité de données (94% de données) sont dans le premier groupe, un groupe de 4,4%, la troisième groupe a 1,6% de données.



FIGURE 30 – Cluster

En utilisant la fonction 'plot3d' fourni par le package 'rgl', nous pouvons visualiser les données en trois dimensions, et nous pouvons visualiser le résultat de clustering.

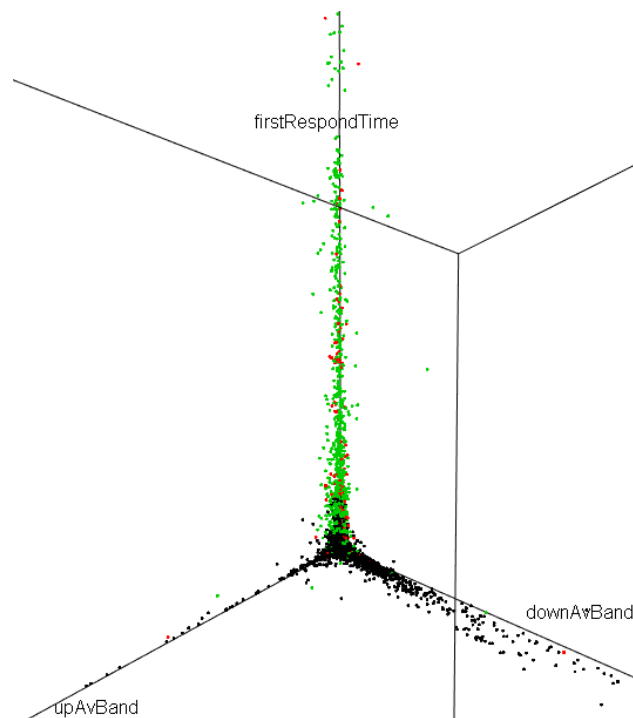


FIGURE 31 – up link Average Band, down link Average Band, first Response Time.

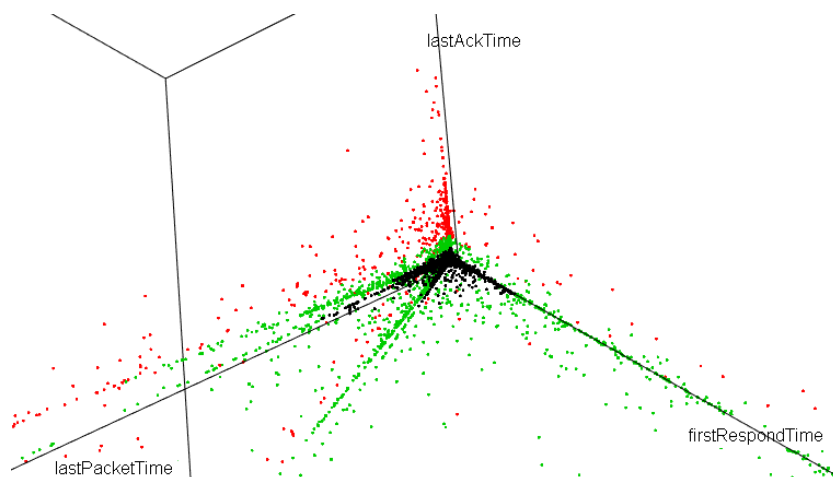


FIGURE 32 – first Respond Time, last Packet Time, last Ack Time.

Nous avons trouver que le données de 'cluster 1' sont les points sont colorent en noir, 'cluster 2' en rouge, 'cluster 3' en vert.

Après le regroupement, nous pouvons utiliser l'algorithme 'k plus proches voisins' pour déterminer le nouveau données font partie de quel groupe .

### 6.5.3 Le Règle d'association

En utilisant le résultat de clustering, nous pouvons transformer les données numérique à données caractère.

Nous y parvenons en quatre étapes

1. D'abord, nous avons trouvé les valeur minimal et maximal de chaque attribut<sup>33</sup>;
2. Ensuite, nous avons trouvé les intervalles de chaque attribut<sup>34</sup>;
3. Puis, nous avons transformé les données numérique à données caractère<sup>35</sup>;
4. À la fin, nous avons utilisé l'algorithme Apriori pour trouver des règles<sup>36</sup>.

	row.names	V1	V2	V3	V4	V5
1	minV	0.06616667	0.0000	0	0	0
2	maxV	170.16000000	1316.0000	8590	8610	7884
3	minV	0.10936691	0.1290	0	3	17902
4	maxV	40.90133333	1004.1145	93496	83559	1220150
5	minV	0.12884211	0.0800	0	4	0
6	maxV	30.15066667	243.9792	150695	92081	54999

FIGURE 33 – trouver les valeur minimal et maximal de chaque attribut

	row.names	V1	V2	V3	V4	V5
1	seuil	0.08776679	0.0400	4295.0	1.5	3942.0
2	seuil	0.11910451	0.1045	51043.0	3.5	12893.0
3	seuil	15.13975439	122.0541	122095.5	4307.0	36450.5
4	seuil	35.52600000	624.0468	0.0	46084.5	637574.5
5	seuil	105.53066667	1160.0573	0.0	87820.0	0.0

FIGURE 34 – défini les intervalles

	row.names	upAvBand	downAvBand	firstRespondTime	lastPacketTime	lastAckTime
1	1	3	3	2	3	1
2	2	3	4	2	3	5
3	3	3	4	1	1	1
4	4	2	4	1	1	1
5	5	3	4	1	1	1
6	6	3	4	1	2	1
7	7	3	4	1	3	1
8	8	3	3	1	1	1
9	10	3	3	1	1	1
10	11	3	3	4	5	1
11	12	3	3	1	1	1

FIGURE 35 – transformer les données numérique à données caractère

inspect(aRule)					
##	lhs	rhs	support	confidence	lift
## 1	{}	=> {lastAckTime=1}	0.8904	0.8904	1.0000
## 2	{}	=> {firstRespondTime=1}	0.8984	0.8984	1.0000
## 3	{}	=> {upAvBand=3}	0.9146	0.9146	1.0000
## 4	{lastPacketTime=2}	=> {firstRespondTime=1}	0.1069	0.9091	1.0119
## 5	{lastPacketTime=2}	=> {upAvBand=3}	0.1061	0.9017	0.9859
## 6	{downAvBand=4}	=> {lastAckTime=1}	0.2660	0.8990	1.0096
## 7	{downAvBand=4}	=> {firstRespondTime=1}	0.2701	0.9128	1.0161
## 8	{downAvBand=4}	=> {upAvBand=3}	0.2748	0.9288	1.0155
## 9	{downAvBand=3}	=> {lastAckTime=1}	0.4929	0.8673	0.9740
## 10	{downAvBand=3}	=> {firstRespondTime=1}	0.4968	0.8743	0.9732
## 11	{downAvBand=3}	=> {upAvBand=3}	0.5180	0.9115	0.9966
## 12	{lastPacketTime=1}	=> {lastAckTime=1}	0.7096	1.0000	1.1231
## 13	{lastPacketTime=1}	=> {firstRespondTime=1}	0.7096	1.0000	1.1131
## 14	{lastPacketTime=1}	=> {upAvBand=3}	0.6551	0.9231	1.0094
## 15	{lastAckTime=1}	=> {firstRespondTime=1}	0.8293	0.9313	1.0367
## 16	{firstRespondTime=1}	=> {lastAckTime=1}	0.8293	0.9231	1.0367
## 17	{lastAckTime=1}	=> {upAvBand=3}	0.8166	0.9172	1.0028
## 18	{upAvBand=3}	=> {lastAckTime=1}	0.8166	0.8929	1.0028
## 19	{firstRespondTime=1}	=> {upAvBand=3}	0.8243	0.9175	1.0032
## 20	{upAvBand=3}	=> {firstRespondTime=1}	0.8243	0.9012	1.0032
## 21	{downAvBand=4,				
##	lastPacketTime=1}	=> {lastAckTime=1}	0.2092	1.0000	1.1231
## 22	{downAvBand=4,				
##	lastPacketTime=1}	=> {firstRespondTime=1}	0.2092	1.0000	1.1131

FIGURE 36 – Les règles été trouvé par l'algorithme règle d'association

## 6.6 Le KmMST

Par clustering le données en beaucoup de partitions, et utilise la technique de MST, nous pouvons surmonter les défaut de K-Means et clustering le données..

### 6.6.1 L'étape de l'algorithme KmMST

1. Regroupement le données ;

La valeur de  $K$  égal à  $n^r$ , le  $n$  égal à la quantité de données,  $r$  est dans l'intervalle de 0 et 1, par défaut,  $r$  égal à 0,5.

En premier, nous avons regroupé les données en K partition.

2. Établir la matrice de distance ;

Après, nous pouvons utiliser les valeurs du centre de chaque partition pour calculer la matrice de distance

	weight	from	to
1	6110.7599	1	2
2	6110.7599	1	3
3	6110.7562	1	4
4	6110.7605	1	5
5	5439.0350	1	6
6	48387.2799	1	7
7	5579.4425	1	8
8	6111.3178	1	9
9	5809.5372	1	10
10	10534.8457	1	11

FIGURE 37 – la matrice de distance

3. Utilise la technique de MST ;

À aide de la package 'igraph' nous pouvons trouver l'arbre couvrant de poids minimal.

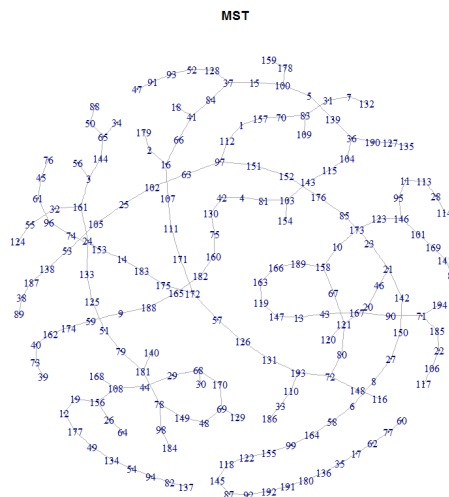


FIGURE 38 – L'arbre couvrant de poids minimal



4. Coupe le arête le plus long

Techniquement, nous pouvons trouver les partitions par couper les arête long, et nous avons tester avec différentes valeurs, mais les sommets sont devenu le point isolé.

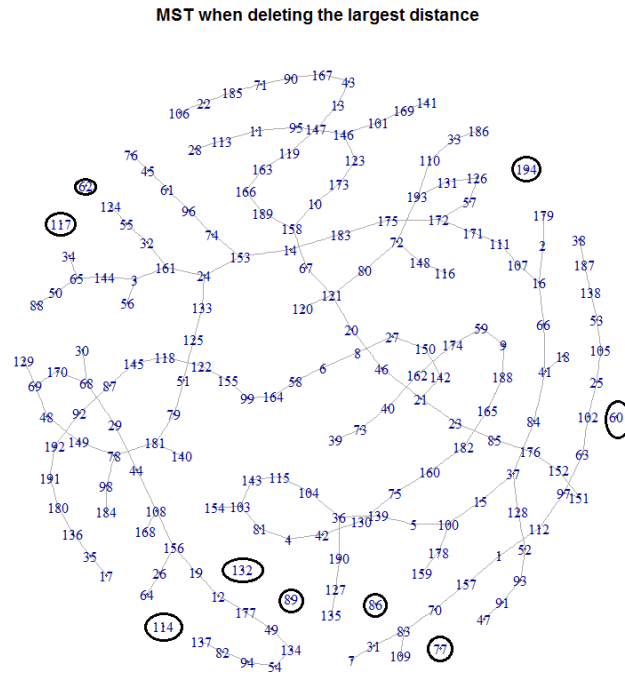


FIGURE 39 – L'arbre couvrant de poids minimal, quand  $C = 10$

## Conclusion

Pendant ces 4 mois de stage, nous avons lu beaucoup de articles, nous avons préparé des dossier et fait des rapport pour le CMCC, aussi nous avons négocié avec les gens du CMCC et les gens du fournisseur de l'équipement. Nous avons perdu beaucoup de temps en négocier et attendre, mais finalement, nous avons trouvé 2 méthode pour résoudre le problème du CMCC.

Dans le rapport, j'ai présenté les algorithmes et la mise en œuvre, mais malheureusement, les résultats ne sont pas satisfaisant, nous avons essayé différent paramètre, différente technique, différent logiciel, mais nous n'avons pas trouvé un résultat satisfaisant. Donc nous croyons que cela est causée du fait que la qualité de données est mal. Et parce que nous avons seulement le données erronées, nous ne pouvons pas évaluer notre techniques.

Cependant, pour l'entreprise CMCC, nous avons fait une démonstration en comment utilise les technique de fouille de données avec ses données, et les inconvénients de son système. Et aussi nous montrons que si il a le données non erronées, comme il peut trouver les informations qu'il a besoin.

Pour moi, j'ai utilisé le langage R et une variété d'algorithme pour résoudre la demande du CMCC. J'ai installé l'Hadoop dans mon ordinateur et géré l'Hadoop en Rstudio. et pendant ce stage, J'ai rencontré des bons amis dans le laboratoire, et acquis des expérience professionnel.

## Les travaux futurs

En attendant de nouvel données, nous devons essayé utiliser l'Hadoop en R, et en même temps essayé optimiser l'algorithme.

## Références

- [1] R&D DÉPARTEMENT DE CMCC. *Interface Specification of China Mobile Signaling Monitoring System(LTE Signal Collection Gateway Part)*.
- [2] Jianhua DU, Shiwen LU, Fangfeng ZHANG *Research of KQI Development Methodology in SQM*,2008.
- [3] Luning ZHAO, Zhuo SUN, Wenbo WANG *Mobile streaming QoE index system and quantify*,2012.
- [4] Rui WANG, Fei SU, Zhengdong HAN, Zilong CAI. *The Recessive Problem Mining and Optimization Research of Voice Service Based on User Behaviors*. édition, 2013.
- [5] Lianjiang ZHU, Bingxian MA, Xuequan ZHAO. *Clustering validity analysis based on silhouette coefficient*, 2010.
- [6] Hongyan WANG, Daiwen WU. *Discussion on digging algorithm of correlation rule for numerical attribute*, 2012.
- [7] Hao OUYANG, Bo CHEN, Zhenjin HUANG, Meng WANG, Zhiwen WANG. *MST Clustering Algorithm Based on K-Means*. 2014.